Statistical Methods for Imaging System Design

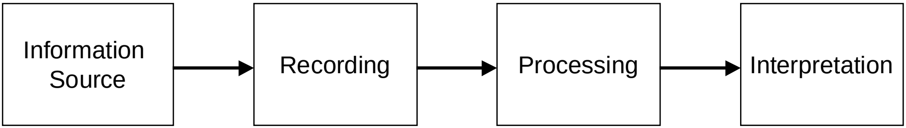
Neal H. Clinthorne,

Division of Nuclear Medicine
The University of Michigan
Ann Arbor, MI

Overview

- The likelihood function or its moments: the forward problem.
- The reverse problem: detection, estimation, figures-of-merit.
- Break
- Examples

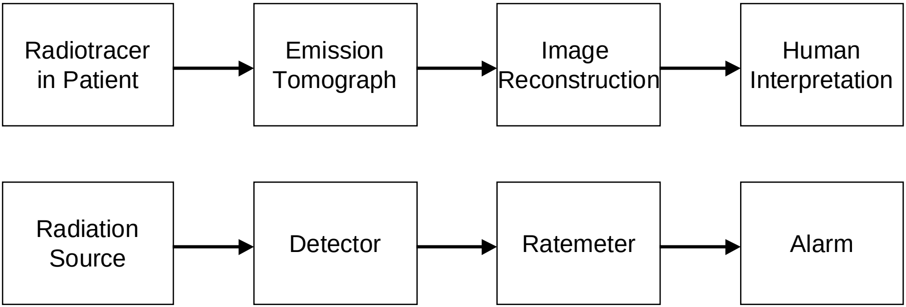
The General Picture



In general the imaging system can be viewed as a chain of blocks. The basic goal is to maximize the transfer of <u>useful</u> information to the interpreter, who then takes an appropriate action.

The block diagram above is oversimplified in the sense that it does not include the feedback paths from the interpreter back to each block.

More Concrete Examples

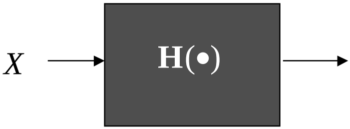


Assuming each system has a random nature, the overall relation of the output or action can be represented by a likelihood function.

The Likelihood Function

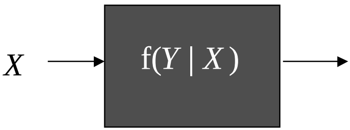
- ***The likelihood function should be the holy grail of the imaging system engineer***
- ***Quantifies randomness in the measurements for a given input***
- ***Can be used to quantify imaging system performance in accomplishing a given task***
- ***Can be used to optimize imaging system performance for a given task***
- ***Is integral to recovering the desired information from the measurements***
- ***Never know the true likelihood, but we attempt to <u>model</u> it***

System Representations



$$Y = \mathbf{H}(X)$$

Very often we only see deterministic blocks in system diagrams



$$Y$$

$$\bar{Y}(X) = \mathrm{E}(Y \mid X) = \int_{-\infty}^{\infty} Y\, \mathrm{f}(Y \mid X)\, dY$$

In this presentation, more attention will be given to blocks in which the observations are only statistically related to the input.

The relationship in the second box above is represented as a conditional probability density function (pdf) where the output, hopefully, bears some relationship to the input. It could just as easily be a conditional probability mass function or even the distribution function.

Examples

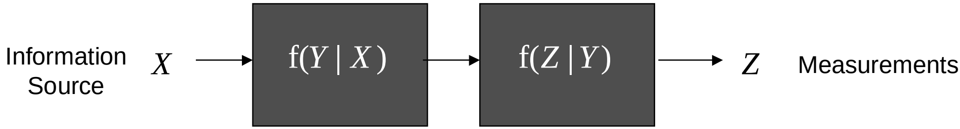
Deterministic Blocks

- Image reconstruction (hopefully)
- Digital processing (almost)

Random Blocks

- Radiotracer injected into patient
- SPECT or PET cameras
- Scintillation cameras
- Radiation detectors
- etc

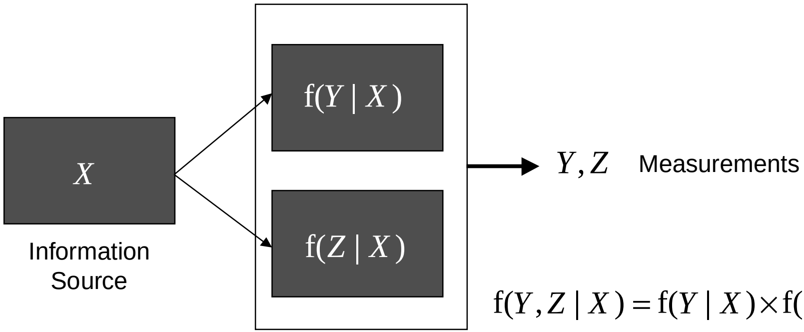
Cascade of Blocks



$$f(Z \mid X) = \int f(Z \mid Y)\, f(Y \mid X)\, dY$$
$$= E\left[f(Z \mid Y) \mid X \right]$$

The desired likelihood describing the statistics of the measurements as a function of the unknown parameters of the information source is obtained by averaging the conditional pdf in the second block over that in the first. These conditional expectations will prove central for determining the overall likelihood function as well as its moments. Note that the expectation operator can be *iterated* to give the likelihood function for any number of blocks.

Parallel Blocks



$$f(Y, Z \mid X) = f(Y \mid X) \times f(Z \mid X)$$

If the measurements are *conditionally independent* given the unknown parameters, the joint-likelihood function assumes a particularly simple form. A common example is the Poisson likelihood function used to model PET and SPECT systems. If the measurements are not conditionally independent, things get complicated very quickly. A common trick to simplify the likelihood in this case is to condition on enough parameters such that the measurements are independent. The unknown values of these “nuisance” parameters are estimated simultaneously with the desired parameters.

Poisson Likelihood

$$\mathrm{P}\left[N_T \mid \Lambda_T\right] = \frac{\mathrm{e}^{-\Lambda_T}\,\Lambda_T^{N_T}}{N_T!}$$

N_T Number of events in interval $[0,T]$

Λ_T Integrated rate over same interval

$$\mathrm{E}\left[N_T \mid \Lambda_T\right] = \Lambda_T$$

$$\sigma^2\left(N_T \mid \Lambda_T\right) = \Lambda_T$$

- Sum of Poisson random variables is again Poisson distributed with mean and variance equal to the sum of the rates or intensities. (Tomography)
- A thinned Poisson process in which events are randomly deleted with probability 1-P is again Poisson distributed with new rate = P X old rate. (Photoelectron emission in a phototube with quantum efficiency P).
- In a Poisson point process occurrence times, spatial locations, or some other parameter may be recorded in addition to the count. If these points are randomly translated (in space, for example), the resulting point process is again Poisson but with a rate function that has been “smeared” by the uncertainty. (Effect of detector resolution in a PET or SPECT system).

Refer to *Random Point Processes* by DL Snyder and MI Miller

Gaussian Likelihood

$$\mathrm{f}\left(y \mid \theta\right) = C \times \mathrm{e}^{-\frac{(y - \bar{y}(\theta))^2}{2\sigma^2}}$$

$\bar{y}(\theta)$ Mean as a function of θ

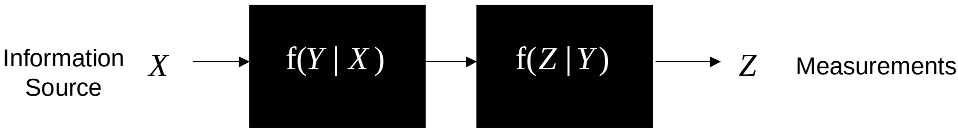
σ^2 Variance

- Completely characterized by its first two moments
- Sum of independent Gaussian distributed random variables is again Gaussian. The output of a cascade of blocks if the means are *linear* functions of the parameters is again Gaussian distributed.
- Limiting probability distribution in many cases

Moments

Often the likelihood function can be difficult to deal with directly. It is usually relatively straightforward to either calculate or approximate its moments.

The two most significant moments are the mean and variance (or covariance).

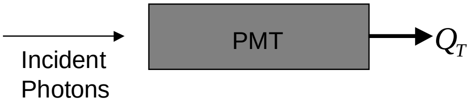


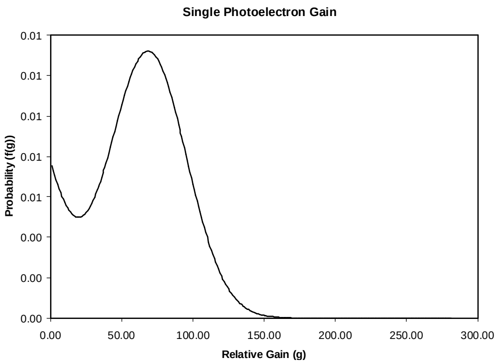
$$\mathrm{E}(Z \mid X) = \iint Z\,\mathrm{f}(Z \mid Y)\,\mathrm{f}(Y \mid X)\,dY dZ = \mathrm{E}_Y\left[\mathrm{E}(Z \mid Y) \mid X\right]$$

$$\sigma^2(Z \mid X) = \mathrm{E}_Y\left[\mathrm{E}(Z^2 \mid Y) \mid X\right] - \mathrm{E}_Y^2\left[\mathrm{E}(Z \mid Y) \mid X\right]$$

Iterated expectations are again handy, although moment generating functions can be used for the sake of confusion.

Example 1: Charge produced by PMT





η - quantum efficiency

$\mathrm{f}(g)$ - pdf of single p.e. gain

$$Q_T = \sum_{i=1}^{N_T} g_i \quad \text{- model for charge}$$

Determine mean and variance of output charge.

Example 1 (cont'd)

Mean:

$$\mathrm{E}[Q_T \mid \Lambda_T] = \mathrm{E}\left[\mathrm{E}\left[\sum_{\mathrm{i}=1}^{N_T} g_i \mid N_T\right] \mid \Lambda_T\right] = \mathrm{E}[N_T \mid \Lambda_T]\overline{g} = \eta\Lambda_T\overline{g}$$

Variance:

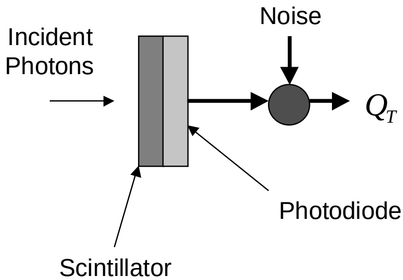
$$\mathrm{E}\left[Q_T^2 \mid \Lambda_T\right] - \mathrm{E}^2[Q_T \mid \Lambda_T] = \mathrm{E}\left[\mathrm{E}\left[\sum_{i=1}^{N_T}\sum_{j=1}^{N_T} g_i g_j \mid N_T\right] \mid \Lambda_T\right] - \eta^2\Lambda_T^2\overline{g}^2$$

$$= \mathrm{E}\left[(N_T^2 - N_T)\overline{g}^2 + N_T\mathrm{E}(\mathrm{g}^2)\right] - \eta^2\Lambda_T^2\overline{g}^2$$

$$= \eta^2\Lambda_T^2\overline{g}^2 + (\eta\Lambda_T - \eta\Lambda_T)\overline{g}^2 - \eta^2\Lambda_T^2\overline{g}^2 + \eta\Lambda_T E(g^2)$$

$$= \eta\Lambda_T E(g^2) = \eta\Lambda_T\overline{g}^2 + \eta\Lambda_T\sigma_g^2$$

Example 2: Charge collected by photodiode



$$Q_T = \sum_{i=1}^{N_T} g_i + n$$

$$\mathrm{E}[\mathrm{n}] = \overline{Q}_{dark}$$

$$\mathrm{E}[Q_T \mid \Lambda_T] = \eta \Lambda_T \overline{g} + \overline{Q}_{dark}$$

$$\sigma_{Q_T}^2 = \eta \Lambda_T \mathrm{E}(g^2) + \sigma_n^2$$

Although the physical device is quite different from the previous, the same model structure can be used with the addition of an additive noise term.

Multiple Dimensions

Unknown Parameters $\theta = [\theta_1, \ldots, \theta_M]^{\mathrm{T}}$

Measurements $\mathbf{y} = [y_1, \ldots, y_N]^{\mathrm{T}}$

Mean measurements $\overline{\mathbf{y}}(\theta) = \mathrm{E}[\mathbf{y} \mid \theta] = [\overline{y}_1(\theta), \ldots, \overline{y}_N(\theta)]^{\mathrm{T}}$

Covariance Matrix $\mathbf{K}_{\mathbf{y}} = \mathrm{E}[\mathbf{y}\mathbf{y}^{\mathrm{T}} \mid \theta] - \mathrm{E}[\mathbf{y} \mid \theta]\mathrm{E}[\mathbf{y}^{\mathrm{T}} \mid \theta]$

Multidimensional Gaussian log - likelihood

$$\log \mathrm{f}(\mathbf{y} \mid \theta) = C + \left[-\frac{1}{2}(\mathbf{y} - \overline{\mathbf{y}}(\theta))^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} (\mathbf{y} - \overline{\mathbf{y}}(\theta)) \right]$$

We are almost always interested in a multidimensional measurement space. However, the number of parameters may range from one to hundreds of thousands (image reconstruction)

Multidimensional Poisson Likelihood

Unknown Parameters $\theta = [\theta_1, \ldots, \theta_M]^{\mathrm{T}}$

Measurements $\mathbf{y} = [y_1, \ldots, y_N]^{\mathrm{T}}$

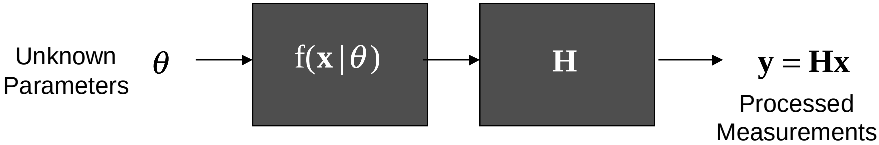
Mean measurements $\bar{\mathbf{y}}(\theta) = \mathrm{E}[\mathbf{y} \mid \theta] = [\bar{y}_1(\theta), \ldots, \bar{y}_N(\theta)]^{\mathrm{T}}$

Covariance Matrix $\mathbf{K}_{\mathbf{y}} = diag\{\bar{y}_i(\theta)\}$

Multidimensional Poisson log - likelihood for count data

$$\log \mathrm{f}(\mathbf{y} \mid \theta) = \sum_{i=1}^{N} [y_i \log \bar{y}_i(\theta) - \bar{y}_i(\theta)] + C$$

Propagating Moments for Linear Processing



Mean : $$\mathrm{E}\left[\mathbf{y}\,|\,\theta\right] = \mathrm{E}\left[\mathbf{H}\mathbf{x}\,|\,\theta\right] = \mathbf{H}\overline{\mathbf{x}}(\theta)$$

Covariance : $$\mathbf{K}_{\mathbf{y}} = \mathrm{E}\left[\mathbf{H}(\mathbf{x}-\overline{\mathbf{x}}(\theta))(\mathbf{x}-\overline{\mathbf{x}}(\theta))^{\mathrm{T}}\mathbf{H}^{\mathrm{T}}\,|\,\theta\right]$$
$$= \mathbf{H}\mathbf{K}_{\mathbf{x}}\mathbf{H}^{\mathrm{T}}$$

For linear processing (matrix multiplication, integral operators, etc.), calculation of moments is straightforward and exact.

Non-linear Processing

$$\mathbf{y} = \mathbf{H}(\mathbf{x})$$

$$\mathbf{y} = \mathbf{H}(\overline{\mathbf{x}}) + \nabla\mathbf{H}(\overline{\mathbf{x}})(\mathbf{x} - \overline{\mathbf{x}}) + \mathrm{H.O.T.}$$

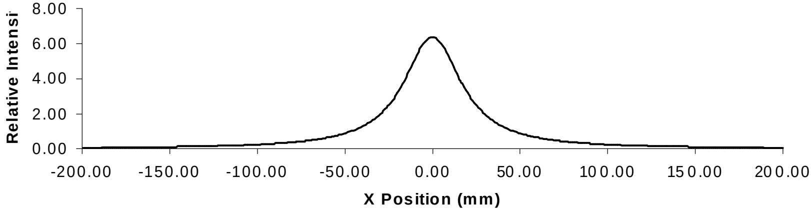
Mean : $\mathrm{E}\left[\mathbf{y} \,|\, \theta\right] \approx \mathbf{H}(\overline{\mathbf{x}}(\theta))$

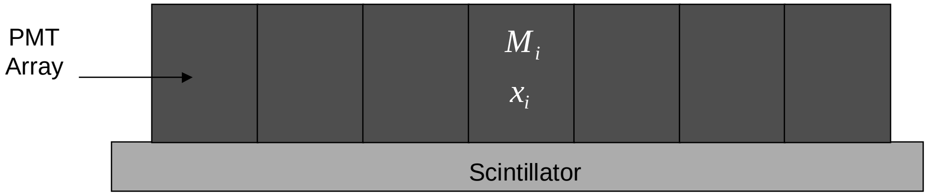
Covariance : $\mathbf{K}_{\mathbf{y}} \approx \nabla\mathbf{HK}_{\mathbf{x}}(\nabla\mathbf{H})^{\mathrm{T}}$

$$\nabla\mathbf{H} = \begin{bmatrix} \dfrac{\partial H_1}{\partial x_1} & \cdots & \dfrac{\partial H_1}{\partial x_N} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial H_M}{\partial x_1} & \cdots & \dfrac{\partial H_M}{\partial x_N} \end{bmatrix}$$

The well-known error propagation formula (here extended to multiple dimensions) is based on a linearization of the non-linearity.

Scintillation Camera Resolution





Centroid Position Estimator

$$\hat{x} = \frac{\sum_i M_i x_i}{\sum_i M_i}$$

Variance of Centroid Estimator

$$\hat{x} = \frac{\sum_i m_i x_i}{\sum_i m_i}$$

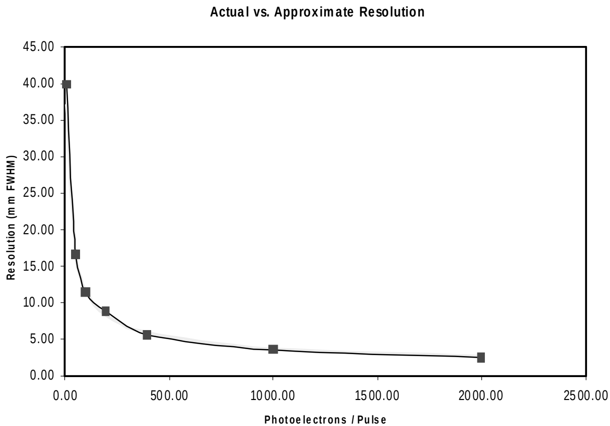
$$Cov(\mathbf{m}) = diag\{\overline{m}_i\}$$

$$\frac{\partial \hat{x}}{\partial m_i} = \frac{1}{\sum_i m_i}\left[x_i - \frac{\sum_i m_i x_i}{\sum_i m_i}\right]$$

$$\frac{m_i}{\sum_i m_i} \approx P_i \Rightarrow \sigma_{\hat{x}}^2 \approx \frac{\sigma_{LSF}^2}{\mathrm{E}\left[\sum_i m_i\right]}$$

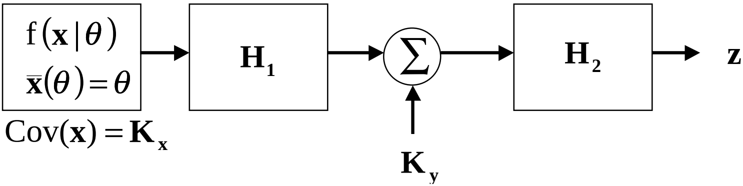
In this case, the resolution is related to the variance of the shape of the light spread function divided by the expected number of scintillation photons in the pulse. This relationship does not hold if the measurement covariance has additional additive noise.

But is the approximation any good?



Filled squares represent Monte Carlo estimates. Solid line is approximation vs. mean scintillation intensity.

Cascade of Linear Blocks with Additive Noise



$$E[\mathbf{z}|\theta]=\mathbf{H}_2\mathbf{H}_1\theta$$ Assuming noise is zero mean

$$\mathbf{K}_\mathbf{z}=\mathbf{H}_2\left(\mathbf{H}_1\mathbf{K}_\mathbf{x}\mathbf{H}_1^T+\mathbf{K}_\mathbf{y}\right)\mathbf{H}_2^T$$

Signal-to-Noise Ratio
Fisher Information

Define “signal” as the change in the measurements for a given change in the parameters, i.e.,

$$\text{Signal} = \nabla \bar{\mathbf{y}}(\theta)$$

Define “noise” by the covariance matrix or autocovariance function:

$$\text{Noise} = \mathbf{K}_{\mathbf{y}}(\theta)$$

Then the Signal-to-Noise ratio, which is also the Fisher Information Matrix for Gaussian and Poisson conditional pdfs is given by:

$$\mathbf{F}_{\theta} = \left(\nabla \bar{\mathbf{y}}(\theta)\right)^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1}(\theta) \nabla \bar{\mathbf{y}}(\theta)$$

For the previous example, the SNR is:

$$\mathbf{F}_{\theta} = \mathbf{H}_1^{\mathrm{T}} \mathbf{H}_2^{\mathrm{T}} \left[\mathbf{H}_2 \left(\mathbf{H}_1 \mathbf{K}_{\mathbf{x}} \mathbf{H}_1^{\mathrm{T}} + \mathbf{K}_{\mathbf{y}}\right) \mathbf{H}_2^{\mathrm{T}}\right]^{-1} \mathbf{H}_2 \mathbf{H}_1$$

Note that processing cannot increase the SNR. At best it stays the same.

Relation to "Classical" System Analysis

In classical imaging system analysis, the processing (or degradation) is assumed linear and shift-invariant. Assuming that the autocovariance is also shift-invariant allows diagonalization of both signal and noise via the Fourier transform

$$\mathbf{H}^{\mathrm{T}}\mathbf{K}_{\mathbf{x}}^{-1}\mathbf{H} = \mathbf{F}\left(\mathbf{F}^{*}\mathbf{H}^{\mathrm{T}}\mathbf{F}\right)\left(\mathbf{F}^{*}\mathbf{K}_{\mathbf{x}}^{-1}\mathbf{F}\right)\left(\mathbf{F}^{*}\mathbf{H}\mathbf{F}\right)\mathbf{F}^{*}$$

$$= \mathbf{F}\left[\frac{\left|\tilde{H}(f)\right|^{2}}{K_{\mathbf{x}}(f)}\right]\mathbf{F}^{*} = \mathbf{F}\left[\frac{MTF^{2}(f)}{NPS(f)}\right]\mathbf{F}^{*}$$

For the previous example, if the Fourier transform diagonalizes it:

$$SNR = \mathbf{F}\left[\frac{\left|\tilde{H}_{1}(f)\right|^{2}\left|\tilde{H}_{2}(f)\right|^{2}}{\left|\tilde{H}_{2}(f)\right|^{2}K_{\mathbf{x}}(f)+K_{\mathbf{y}}(f)}\right]\mathbf{F}^{*}$$

DQE and all that

Detective Quantum Efficiency can be thought of as a transfer function for SNRs

$$DQE = \frac{SNR_{out}}{SNR_{in}}$$

For the previous example:

$$DQE(f) = \left[\frac{\left|\tilde{H}_1(f)\right|^2 \left|\tilde{H}_2(f)\right|^2}{\left|\tilde{H}_2(f)\right|^2 K_{\mathbf{x}}(f) + K_{\mathbf{y}}(f)} \right] \times \left[\frac{K_{\mathbf{x}}(f)}{\left|\tilde{H}_1(f)\right|^2} \right]$$

$$= \left[\frac{\left|\tilde{H}_2(f)\right|^2}{\left|\tilde{H}_2(f)\right|^2 + \dfrac{K_{\mathbf{y}}(f)}{K_{\mathbf{x}}(f)}} \right]$$

While DQE can be defined the same way for spatially-varying systems, it is not often used.

Part I Summary

- *The likelihood function can be used to represent how the measurements in a complex system are related to underlying parameters of interest.*
- *The overall likelihood function for a system can be obtained using iterated expectations.*
- *Moments of the likelihood function can be calculated in a similar fashion*

- *Did not mention the important topics of how to construct models for individual boxes or model validation*
- *Did not yet mention system optimization using the likelihood*

Part II: The inverse problem

OK, we have the likelihood.

So what?

Why is it useful?

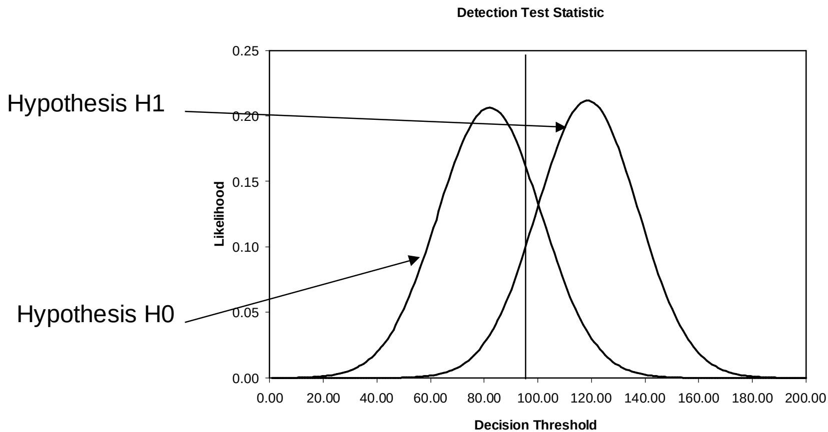
General Tasks

Detection | Classification | Estimation

Sometimes you're wrong | You're wrong more often | You're always wrong. But how wrong?

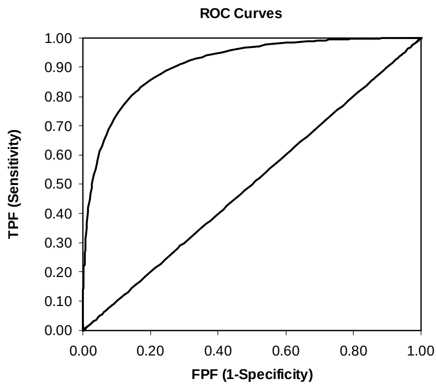
Task types can be broken down into roughly three categories. Detection is generally a binary task that decides whether a signal is present (e.g., a tumor). Classification assigns the observations to one of N signal categories (a PET block detector, for example). Estimation generally results in point estimates of continuous parameters (pixel intensity in image reconstruction).

Detection



The optimum detector for a binary decision task is the likelihood ratio test, which of course depends upon the likelihood functions derived previously. ROC curves can be plotted by sweeping the decision threshold through its range

Receiver Operating Characteristic (ROC) Curve



Ideal Observers for the Signal Known Exactly Binary Hypothesis Test

For the case in which the signal and background are known exactly (SKE / BKE), the optimum detector assumes a convenient form. As in more complex detection problems, the optimum test statistic is the likelihood ratio test (or difference of log-likelihoods); however, it is relatively easy to calculate for Poisson and Gaussian noise.

Step 1: Form the likelihood ratio statistic

$$t(\mathbf{y}) = \mathrm{L}(\mathbf{y} \,|\, \mathrm{H}_1) - \mathrm{L}(\mathbf{y} \,|\, \mathrm{H}_0)$$

$$\mathrm{L}(\mathbf{y} \,|\, \mathrm{H}_\mathrm{i}) = \log \mathrm{f}(\mathbf{y} \,|\, \mathrm{H}_\mathrm{i})$$

Step 2: Compare the test statistic value to a decision threshold. If it is greater choose H1 else choose H0. Can evaluate system performance by plotting

$$\mathrm{f}(t(\mathbf{y}) \,|\, \mathrm{H}_1) \text{ and } \mathrm{f}(t(\mathbf{y}) \,|\, \mathrm{H}_0)$$

Methods developed previously can be used to calculate the requisite pdfs or moments

Ideal Observer: Gaussian Case

If the likelihoods under both hypotheses are Gaussian distributed the LR test statistic is also Gaussian:

$$t(\mathbf{y}) = \mathbf{y}^{\mathrm{T}}\left(\mathbf{K}_1^{-1}\overline{\mathbf{y}}(\mathrm{H}_1) - \mathbf{K}_0^{-1}\overline{\mathbf{y}}(\mathrm{H}_0)\right)$$

$$\mathrm{E}\left[t(\mathbf{y}) \mid \mathrm{H}_{\mathrm{i}}\right] = \left(\overline{\mathbf{y}}(\mathrm{H}_{\mathrm{i}})\right)^{\mathrm{T}}\left(\mathbf{K}_1^{-1}\overline{\mathbf{y}}(\mathrm{H}_1) - \mathbf{K}_0^{-1}\overline{\mathbf{y}}(\mathrm{H}_0)\right)$$

$$\sigma^2\left(t(\mathbf{y}) \mid \mathrm{H}_{\mathrm{i}}\right) = \left(\mathbf{K}_1^{-1}\overline{\mathbf{y}}(\mathrm{H}_1) - \mathbf{K}_0^{-1}\overline{\mathbf{y}}(\mathrm{H}_0)\right)^{\mathrm{T}}\mathbf{K}_{\mathbf{i}}\left(\mathbf{K}_1^{-1}\overline{\mathbf{y}}(\mathrm{H}_1) - \mathbf{K}_0^{-1}\overline{\mathbf{y}}(\mathrm{H}_0)\right)$$

If the covariance is the same for both then more simplification results

$$t(\mathbf{y}) = \mathbf{y}^{\mathrm{T}}\mathbf{K}^{-1/2}\mathbf{K}^{-1/2}\left(\overline{\mathbf{y}}(\mathrm{H}_1) - \overline{\mathbf{y}}(\mathrm{H}_0)\right)$$

The data is “pre-whitened” to remove correlations and then filtered with a modified signal shape (I.e., “Matched Filtering”)

Ideal Observer: Poisson Case

Here, the density of the LR test under either hypothesis is difficult to calculate; however, its moments can readily be computed. Making the assumption that LR test is approximately Gaussian distributed allows calculation of detectabiltiy, ROC curves, etc.

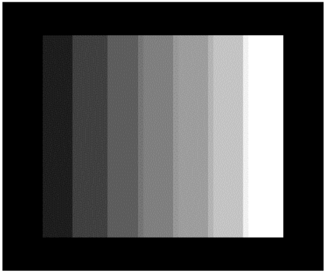
$$t(\mathbf{y}) = \mathbf{y}^{\mathrm{T}}\left(\log \overline{\mathbf{y}}(\mathrm{H}_1) - \log \overline{\mathbf{y}}(\mathrm{H}_0)\right)$$

$$\mathrm{E}\left[t(\mathbf{y}) \,|\, \mathrm{H}_{\mathrm{i}}\right] = \left(\overline{\mathbf{y}}(\mathrm{H}_{\mathrm{i}})\right)^{\mathrm{T}}\left(\log \overline{\mathbf{y}}(\mathrm{H}_1) - \log \overline{\mathbf{y}}(\mathrm{H}_0)\right)$$

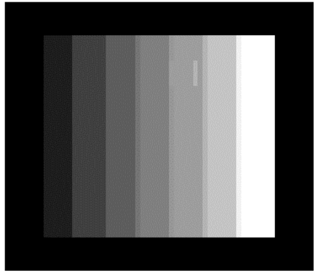
$$\sigma^2\left(t(\mathbf{y}) \,|\, \mathrm{H}_{\mathrm{i}}\right) = \left(\log \overline{\mathbf{y}}(\mathrm{H}_1) - \log \overline{\mathbf{y}}(\mathrm{H}_0)\right)^{\mathrm{T}} diag\left(\overline{\mathbf{y}}(\mathrm{H}_{\mathrm{i}})\right)\left(\log \overline{\mathbf{y}}(\mathrm{H}_1) - \log \overline{\mathbf{y}}(\mathrm{H}_0)\right)$$

Detection Example

The image on the right contains a small hot-spot in addition to the background of the image on the left. We add Poisson noise and empirically estimate the statistics of the LR test.

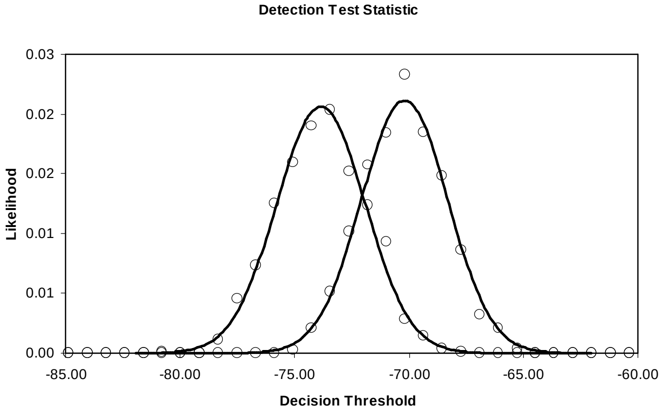


H0



H1

And does the approximation work?



The open circles are histogrammed results of applying 1000 trials of the ideal observer when the true hypothesis is known.

Problems with Simple Detection Tasks

- **SKE binary hypothesis test is too simple. Often leads to useless optimization results.**
- **In imaging, the ideal observer does not match human performance well--especially in correlated noise.**
- **More complex detection problems can become intractable very quickly.**

One thing's for sure: if a system does not perform well for the simple SKE / BKE task, then it's not likely to work better for more complex jobs.

Estimation

Rather than to classify a signal, the goal is to obtain an estimate of the values of unknown parameters as a function of the measurements (the likelihood function will again play a central role).

Define:

$$\hat{\theta}(\mathbf{y}) \quad \text{as an estimate of} \quad \theta$$

In non-Bayesian parameter estimation, the two most important figures-of-merit for an estimator are its *bias*

$$\mathbf{b}_{\hat{\theta}}(\theta) = \mathrm{E}\left[\hat{\theta}(\mathbf{y}) \mid \theta\right] - \theta$$

and its covariance matrix

$$\mathbf{K}_{\hat{\theta}}(\theta) = \mathrm{E}\left[\hat{\theta}(\mathbf{y})\left(\hat{\theta}(\mathbf{y})\right)^{\mathrm{T}} \mid \theta\right] - \mathrm{E}\left[\hat{\theta}(\mathbf{y}) \mid \theta\right]\mathrm{E}\left[\left(\hat{\theta}(\mathbf{y})\right)^{\mathrm{T}} \mid \theta\right]$$

The MSE of the estimate for each parameter is the sum of the variance and the squared bias:

$$MSE\left(\hat{\theta}_i\right) = b_i^2 + K_{ii}$$

the overall MSE matrix is given by

$$\mathbf{K}_{\hat{\theta}} + \mathbf{b}_{\hat{\theta}}\mathbf{b}_{\hat{\theta}}^{\mathrm{T}}$$

More Figures-of-Merit

Actual bias is strongly dependent on, say, the source distribution in tomographic reconstruction. To overcome this problem when approximating system performance one may wish to examine the tradeoff between variance and resolution, the width of the impulse response, or with some aspect of the *bias-gradient*

$$\nabla_{\theta}\mathbf{b}_{\hat{\theta}}(\theta) = \nabla_{\theta}\mathrm{E}\left[\hat{\theta}(\mathbf{y}) \mid \theta\right] - \mathbf{I}$$

A column of the first term on the right hand side is the small-signal impulse response of the estimator for the corresponding element of the parameter vector. A row of the matrix is the mean estimator gradient. Often the two either are the same or are close, but they don't have to be!

In an upcoming slide, we'll use the norm of the bias-gradient as a measure of the width of the response.

Maximum Likelihood Estimation

There are many reasons likelihood based estimators are preferable over other types. The basic idea behind ML estimation is to choose the value of the unknown parameters that maximizes the likelihood of observing the measurements:

$$\hat{\theta} = \arg\max_{\theta} \log \mathrm{f}(\mathbf{y} \mid \theta)$$

Differentiating the log-likelihood, setting it to zero and solving the resulting estimation equations for the unknown parameters gives

$$\nabla^{10} L(\theta, \mathbf{y}) = (\nabla \bar{\mathbf{y}}(\theta))^{\mathrm{T}} \, diag^{-1}(\bar{\mathbf{y}}(\theta))[\mathbf{y} - \bar{\mathbf{y}}(\theta)] = 0$$

for the Poisson case and

$$\nabla^{10} L(\theta, \mathbf{y}) = (\nabla \bar{\mathbf{y}}(\theta))^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} [\mathbf{y} - \bar{\mathbf{y}}(\theta)] = 0$$

for the Gaussian likelihood (“least-squares” estimation).

Generalized Least-Squares

Often the likelihood function is not completely known; however, the first and second moments as a function of the parameters are. In this case, Generalized least-squares estimation is often an alternative. Because the covariance model may vary as a function of the parameters, we specify the GLS estimator through its estimator equations.

$$\nabla^{10} L(\theta, \mathbf{y}) = \left(\nabla \bar{\mathbf{y}}(\theta)\right)^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1}(\theta)\left[\mathbf{y} - \bar{\mathbf{y}}(\theta)\right] = 0$$

Note that this is the same as the least-squares estimator in the previous slide except that the covariance can be a function of the parameters. Although it will not be shown here, GLS estimators have good asymptotic bias and variance properties.

The Cramer-Rao Lower Bound

Surprisingly enough, the covariance matrix of any unbiased estimator can be lower-bounded by a function depending *only* on the log-likelihood. Define first the Fisher Information Matrix as

$$F_{jk} = -\mathrm{E}\left[\frac{\partial^2 \log \mathrm{f}(\mathbf{y} \mid \theta)}{\partial\theta_j \partial\theta_k}\right]$$

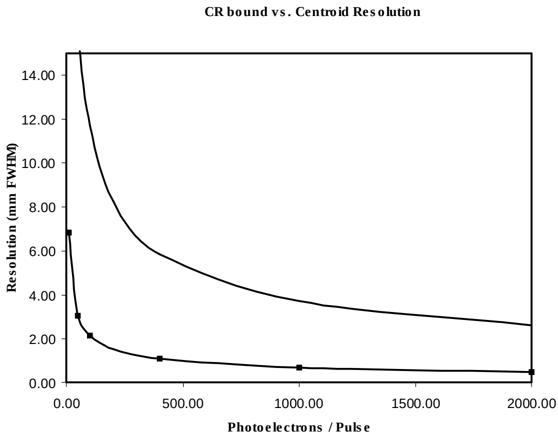
The CR bound lower bounds the covariance matrix of any unbiased estimator in the sense that $\mathbf{K}_{\hat{\theta}} - \mathbf{F}_{\hat{\theta}}^{-1}$ is positive definite.

For both the Gaussian and Poisson likelihoods presented, the Fisher Information takes the following form:

$$\mathbf{F}_{\theta} = \left(\nabla \overline{\mathbf{y}}(\theta)\right)^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1} \nabla \overline{\mathbf{y}}(\theta)$$

$$\mathbf{K}_{\mathbf{y}} = diag\left(\overline{\mathbf{y}}(\theta)\right) \quad \text{for Poisson measurements}$$

Application to Scintillation Camera Example



At the higher photon rates the ML estimator will nearly achieve the CR bound in this case

Yet Another Covariance Approximation

Begin by linearizing the estimator around the mean value of the measurements

$$\hat{\theta}(\mathbf{y}) \approx \hat{\theta}(\overline{\mathbf{y}}(\theta)) + \nabla \hat{\theta}(\overline{\mathbf{y}}(\theta))(\mathbf{y} - \overline{\mathbf{y}}(\theta))$$

$$\mathrm{E}\left[\hat{\theta}(\mathbf{y}) \,|\, \theta\right] \approx \hat{\theta}(\overline{\mathbf{y}}(\theta))$$

$$\mathbf{K}_{\hat{\theta}} \approx \nabla \hat{\theta}(\overline{\mathbf{y}}(\theta)) \mathbf{K}_{\mathbf{y}} \left(\nabla \hat{\theta}(\overline{\mathbf{y}}(\theta))\right)^{\mathrm{T}}$$

Now recognizing that the estimator is an implicit function of the measurements

$$\nabla^{20} L(\theta, \mathbf{y}) \, \nabla_{\mathbf{y}} \hat{\theta}(\mathbf{y}) + \nabla^{11} L(\theta, \mathbf{y}) = 0$$

$$\mathbf{K}_{\hat{\theta}} \approx \left[\nabla^{20} L(\theta, \overline{\mathbf{y}})\right]^{-1} \nabla^{11} L(\theta, \overline{\mathbf{y}}) \mathbf{K}_{\mathbf{y}} \left(\nabla^{11} L(\theta, \overline{\mathbf{y}})\right)^{\mathrm{T}} \left[\nabla^{20} L(\theta, \overline{\mathbf{y}})\right]^{-1}$$

With little more effort, this technique can be used to approximate the covariance of implicitly defined penalized estimators.

Variance of ML Estimator

Now apply the appropriate technique to the Poisson likelihood function:

$$\nabla^{10} L(\theta, \mathbf{y}) = \left(\nabla \overline{\mathbf{y}}(\theta)\right)^{\mathrm{T}} diag^{-1}\left(\overline{\mathbf{y}}(\theta)\right)\left[\mathbf{y} - \overline{\mathbf{y}}(\theta)\right] = 0$$

$$\mathbf{K}_{\hat{\theta}} \approx \left[\nabla^{20} L(\theta, \overline{\mathbf{y}})\right]^{-1} \nabla^{11} L(\theta, \overline{\mathbf{y}}) \mathbf{K}_{\mathbf{y}} \left(\nabla^{11} L(\theta, \overline{\mathbf{y}})\right)^{\mathrm{T}} \left[\nabla^{20} L(\theta, \overline{\mathbf{y}})\right]^{-1}$$

$$\nabla^{20} L(\theta, \overline{\mathbf{y}}) = \left(\nabla \overline{\mathbf{y}}(\theta)\right)^{\mathrm{T}} diag^{-1}\left(\overline{\mathbf{y}}(\theta)\right)\nabla \overline{\mathbf{y}}(\theta) = \mathbf{F}_{\theta}$$

$$\nabla^{11} L(\theta, \overline{\mathbf{y}}) = \left(\nabla \overline{\mathbf{y}}(\theta)\right)^{\mathrm{T}} diag^{-1}\left(\overline{\mathbf{y}}(\theta)\right)$$

$$\mathbf{K}_{\hat{\theta}} \approx \mathbf{F}_{\theta}^{-1}$$

In other words, the approximate covariance of the ML estimate for Poisson (Gaussian too) distributed measurements is given by the CR bound.

Stabilizing the Estimation Problem

More often than not, some form of stabilization has to be applied to the estimator. This is especially true in image reconstruction (as we've seen). There are several methods of stabilization. Virtually all of them work through the action of adding some *bias* to the parameter estimator. The following techniques are commonly used:

- Restriction of allowable estimates through equality constraints on the parameter vector
- Use of inequality constraints, such as non-negativity
- Post-smoothing the estimate
- Bayesian methods can be used if an actual *a priori* pdf of the random parameters is known (exponential decay on depth-of-interaction probability in a scintillator)
- Penalized methods, where certain undesirable properties of the estimates are discouraged in the objective function (many examples in the image reconstruction section)

There are more. Practical estimators often use combinations of these methods.

Covariance of Penalized Estimators

A common Poisson likelihood term augmented with a quadratic penalty on the parameter vector leads to the following estimator equations:

$$\nabla^{10} L(\theta, \mathbf{y}) = \left(\nabla \overline{\mathbf{y}}(\theta)\right)^{\mathrm{T}} diag^{-1}\left(\overline{\mathbf{y}}(\theta)\right)\left[\mathbf{y} - \overline{\mathbf{y}}(\theta)\right] + \beta \mathbf{R} \theta = 0$$

$$\tilde{\theta} = \hat{\theta}\left(\overline{\mathbf{y}}(\theta)\right)$$

$$\mathbf{K}_{\hat{\theta}} \approx \left[\nabla^{20} L(\tilde{\theta}, \overline{\mathbf{y}})\right]^{-1} \nabla^{11} L(\tilde{\theta}, \overline{\mathbf{y}}) \mathbf{K}_{\mathbf{y}}(\theta) \left(\nabla^{11} L(\tilde{\theta}, \overline{\mathbf{y}})\right)^{\mathrm{T}} \left[\nabla^{20} L(\tilde{\theta}, \overline{\mathbf{y}})\right]^{-1}$$

$$\nabla^{20} L(\tilde{\theta}, \overline{\mathbf{y}}) = \left(\nabla \overline{\mathbf{y}}(\tilde{\theta})\right)^{\mathrm{T}} diag^{-1}\left(\overline{\mathbf{y}}(\tilde{\theta})\right) \nabla \overline{\mathbf{y}}(\tilde{\theta}) + \beta \mathbf{R}$$

$$\nabla^{11} L(\tilde{\theta}, \overline{\mathbf{y}}) = \left(\nabla \overline{\mathbf{y}}(\tilde{\theta})\right)^{\mathrm{T}} diag^{-1}\left(\overline{\mathbf{y}}(\tilde{\theta})\right)$$

Note that this is almost the same as the previous derivation with the exception that the expression now depends on the penalty matrix as well as on <u>both</u> true and estimated parameter values

CR Bound for Biased Estimators

Sometimes a small estimator bias will significantly reduce variance. The previous CR bound presented was only for unbiased estimators. Another form of the CR bound is applicable to biased estimators:

$$\mathbf{K}_{\hat{\theta}} \geq \left(\mathbf{I} + \nabla \mathbf{b}_{\hat{\theta}}\right) \mathbf{F}_{\theta}^{-1} \left(\mathbf{I} + \nabla \mathbf{b}_{\hat{\theta}}\right)^{\mathrm{T}}$$

Note that the bound in its nascent form is not extremely useful. It is only a lower bound on the covariance for the class of estimators *having the same bias-gradient*. Not likely a very interesting class of functions.

The Uniform CR Bounds

While the biased bound itself is somewhat restrictive, it serves as a basis for the development of a more useful lower bound for biased estimators. By solving a constrained minimization problem, where the objective is to minimize the covariance over the class of all estimators having a specific *bias-gradient norm,* the Uniform CR Bound is obtained. This form lower bounds the covariance matrix for all estimators with a bias-gradient norm equal to or less than the specified norm.

A particularly useful form of the Uniform CR bound is given by the following pair of parametric equations:

$$\delta_i(\theta,\lambda) = \left\| \left(\mathbf{I} + \lambda \mathbf{F}_\theta\right)^{-1} \mathbf{e}_i \right\| \qquad \text{Bias - gradient norm}$$

$$B_i(\theta,\lambda) = \lambda^2 \mathbf{e}_i^{\mathrm{T}} \left(\mathbf{I} + \lambda \mathbf{F}_\theta\right)^{-1} \mathbf{F}_\theta \left(\mathbf{I} + \lambda \mathbf{F}_\theta\right)^{-1} \mathbf{e}_i$$

$$\mathbf{e}_i = \left[\left\{\delta_{ij}\right\}\right]$$

The bound is calculated by sweeping the parameter λ through the range of 0 to infinity. Although the above formulation uses the Euclidian norm to measure the bias-gradient, other norms can be used with appropriate modification.

What are CR bounds good for?

- ***CR Bounds are excellent tools for indicating when an estimation task is <u>not</u> possible (or practical) given the measurements***

- ***It serves as a performance yardstick for suboptimal estimators. If the estimator already has performance close to the bound, why bother to look further?***

- ***The ML estimator (penalized ML) estimator will asymptotically achieve the CR bound (uniform CR bound).***

- ***Can be used, with a great deal of caution, for optimizing system performance***

And what might they not be good for?

- ***CR Bounds are often overly optimistic. Especially true when either the noise level is high <u>or</u> when the second (or higher) derivatives of the expected measurements w.r.t. the parameters are large.***

- ***While the ML estimator will achieve the bound asymptotically, it may approach the bound only at unrealistically low noise levels. At high noise levels estimator performance can be significantly worse than predicted.***

Calculating Covariance and Other Performance Measures

Most of the covariance expressions have a form similar to

$$\mathbf{K}_{\hat{\theta}} \approx \mathbf{Q}^{-1}\mathbf{Z}\mathbf{Q}^{-1}$$

While for small problems the matrix can be calculated directly, for larger problems such as image reconstruction, obtaining the entire covariance matrix is impractical. In this case, an entire row of the matrix can be computed by solving two successive **Ax=b** type problems. For example, to calculate the covariance corresponding to the i-th parameter:

$$\mathbf{K}_{\hat{\theta}}\mathbf{e}_i \approx \mathbf{Q}^{-1}\mathbf{Z}\mathbf{Q}^{-1}\mathbf{e}_i$$

$$\text{Solve:} \quad \mathbf{e}_i = \mathbf{Q}\mathbf{m}_i$$

$$\mathbf{z}_i = \mathbf{Z}\mathbf{m}_i$$

$$\text{Solve:} \quad \mathbf{z}_i = \mathbf{Q}\mathbf{k}_{i,:}$$

Other performance measures can be calculated in a similar way.

Summary Part II

- ***The likelihood is an integral part of evaluating performance in detection, classification, and estimation tasks***

- ***Detection performance can be quantified via ROC curves and related measures. Ideal observers for Poisson and Gaussian distributed measurements presented for SKE / BKE task.***

- ***Estimation performance quantified by bias and covariance. Approximations presented for penalized and unpenalized estimators. CR lower bounds on covariance.***

Part III: Examples and Applications

Is this stuff good for anything?

Scintillation Cameras

What resolution can be achieved?

How is it affected by PMT size?

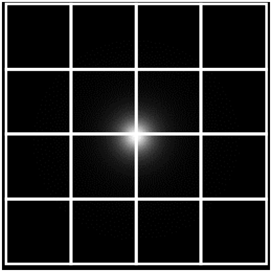
Can depth--of-interaction be estimated?

What's the effect of photodetector noise?

This LSF was used in the following calculations. All scintillator surfaces were black except for the detection surface.

$$LSF(x, y; x_0, y_0, z_0, E) = \frac{1}{2\pi} \frac{E z_0}{\left(z_0^2 + (x - x_0)^2 + (y - y_0)^2\right)^{3/2}} \propto \cos^3 \psi$$

Scintillation Cameras



LSF and PMT Array

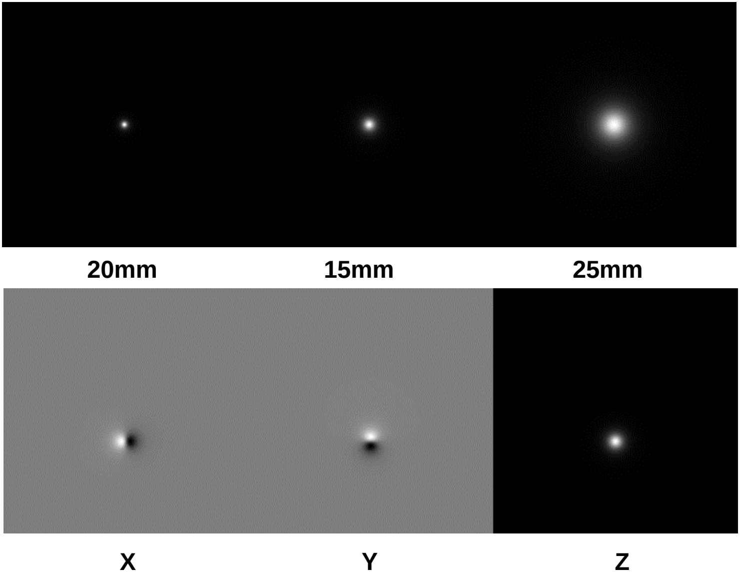
Crystal Size	**40 x 40 cm**
Thickness	**25 mm**
# Photoelectrons	**~400**
Depth-of-Interaction	**10 mm from entrance**

$$\theta = [x_0, y_0, E]$$

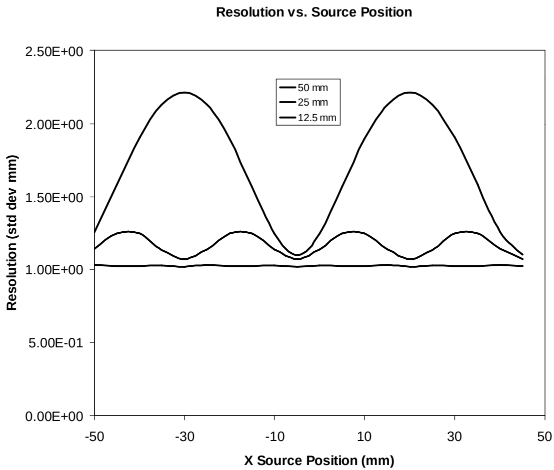
$$\mathbf{K}_{\hat{\theta}} \geq \left[(\nabla \overline{\mathbf{y}}(\theta))^{\mathrm{T}} \, diag^{-1}(\overline{\mathbf{y}}(\theta)) \nabla \overline{\mathbf{y}}(\theta) \right]^{-1}$$

Approximate Spatial and Energy Resolution

LSF and Gradients



Resolution Scan for Various Size PMTs



Estimating Depth-of-Interaction

Shown below are the CR bound matrices for the case where the depth is known (top) and where the depth must be estimated (bottom).

X	***1.14***	**-7.20E-17**	**1.11E-14**
Y	**-7.20E-17**	***1.14***	**-9.42E-16**
ENERGY	**1.11E-14**	**-9.42E-16**	***432***

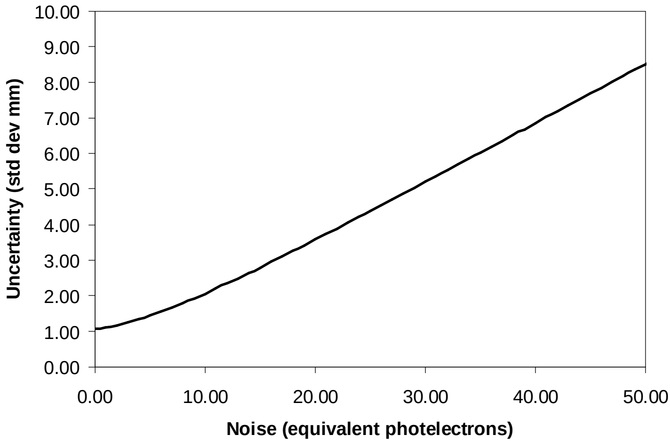
X	***1.14***	**-7.20E-17**	**-1.76E-15**	**1.08E-13**
Y	**-7.20E-17**	***1.14***	**1.25E-16**	**-7.83E-15**
Z	**-1.76E-15**	**1.25E-16**	***1.18***	**-65.2**
ENERGY	**1.08E-13**	**-7.83E-15**	**-65.2**	***4030***

Note the strong correlation between the depth and energy estimates for this LSF!

Effects of Additive Detector Noise

The effects of additive, independent photodetector noise on the resolution can be modeled by adding the detector noise variance to the variance of the "self-noise":

$$\mathbf{K}_{\hat{\theta}} \geq \left[\left(\nabla \overline{\mathbf{y}}(\theta) \right)^{\mathrm{T}} diag^{-1} \left(\overline{\mathbf{y}}(\theta) + \sigma_D^2 \right) \nabla \overline{\mathbf{y}}(\theta) \right]^{-1}$$



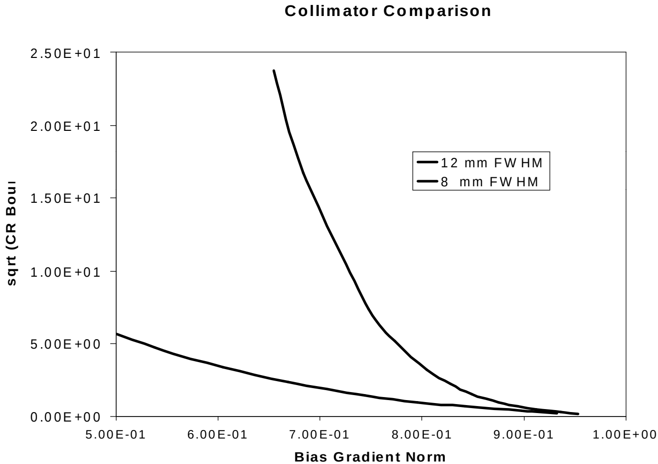
Is It Better to Deconvolve or to use a High Resolution Collimator in the First Place?

In mechanically collimated systems, there's a fundamental tradeoff between resolution and efficiency. Given that the data can be processed, it it better to use a high-efficiency, low-resolution system and then process the data to the desired resolution or to use a high-resolution collimator?

Approach: Model resolution and efficiency of collimators using common expressions. Use Uniform CR bound to compare performance as a function of desired smoothing.

$$FWHM \approx \sqrt{\frac{8\ln 2}{\pi}}\frac{\sqrt{\text{Hole Area}}}{\text{Thickness}} \qquad \text{Efficiency} = F \times \frac{\text{Hole Area}}{4\pi \times \text{Thickness}^2}$$

Uniform CR Bound for Collimators

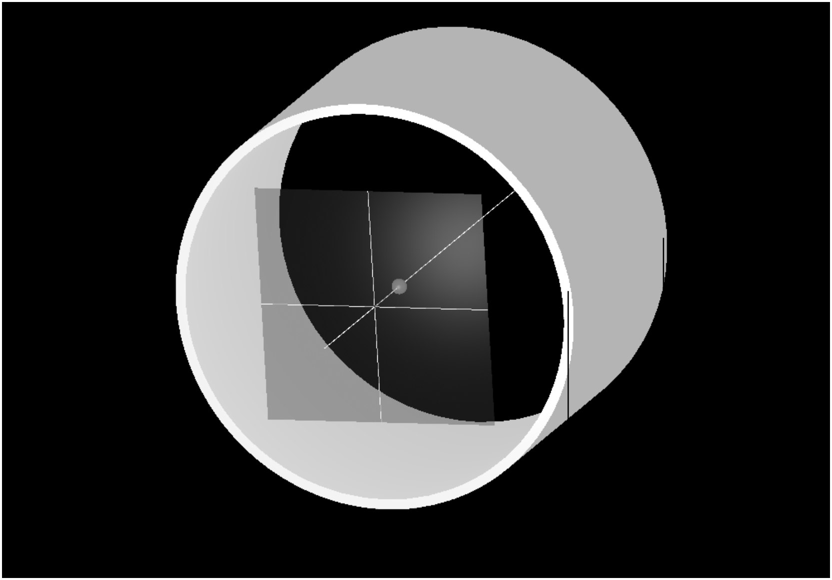


At least for this simple task, the higher resolution collimator should be used.

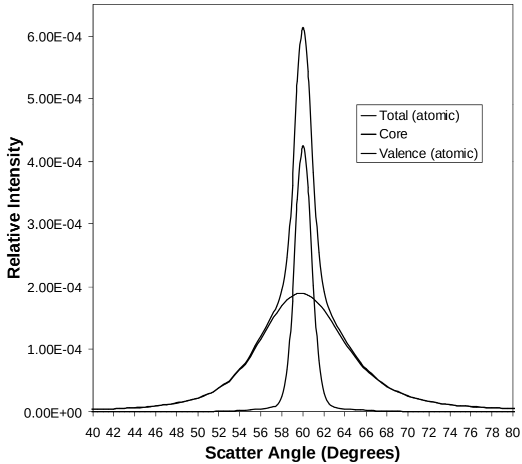
Choosing a Scattering Detector for a Compton-Scatter Camera

Problem: Different scattering materials will exhibit different angular uncertainties due to the effect of Doppler broadening. Which material is best?

Ring Scatter Camera Geometry

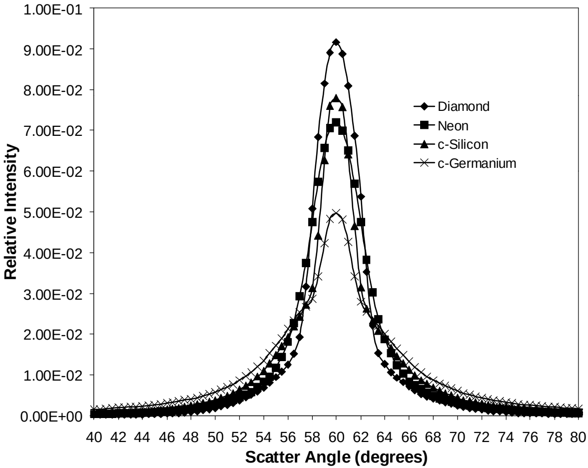


Anatomy of Doppler-Broadened Spectrum



- **The narrow central peak (valence electrons) is important for imaging high spatial frequencies**
- **The broad base (core electrons) degrades performance**

Broadening for Several Materials at 140 keV



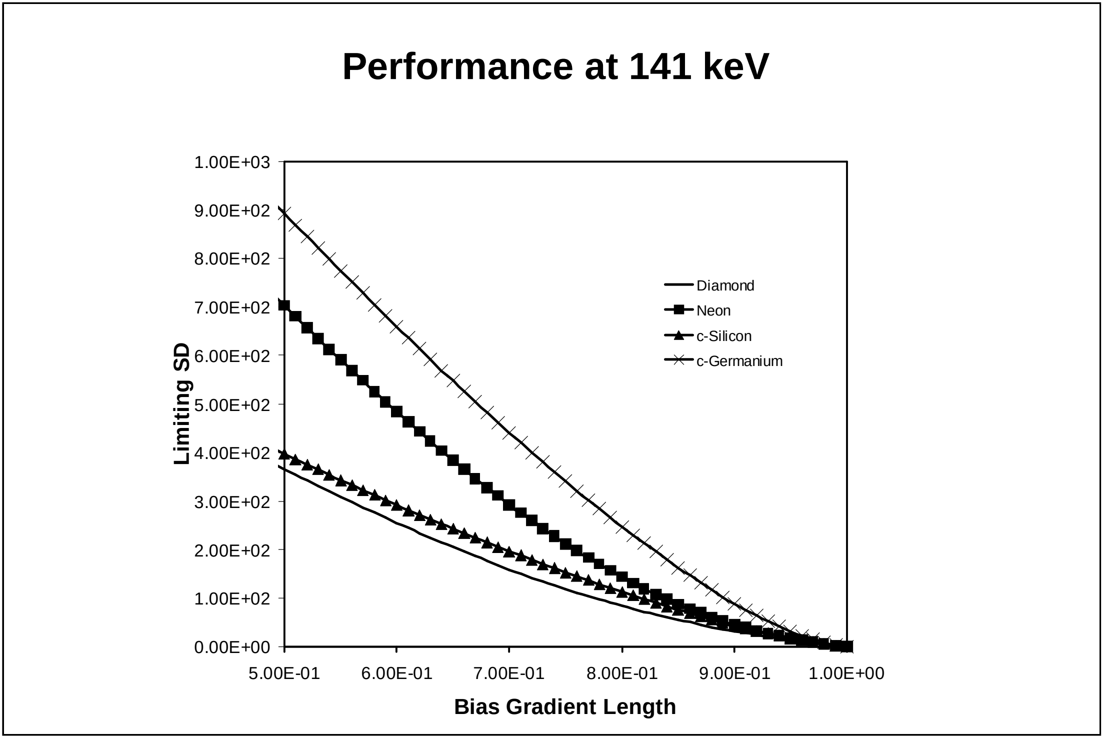
Normalized to equal area

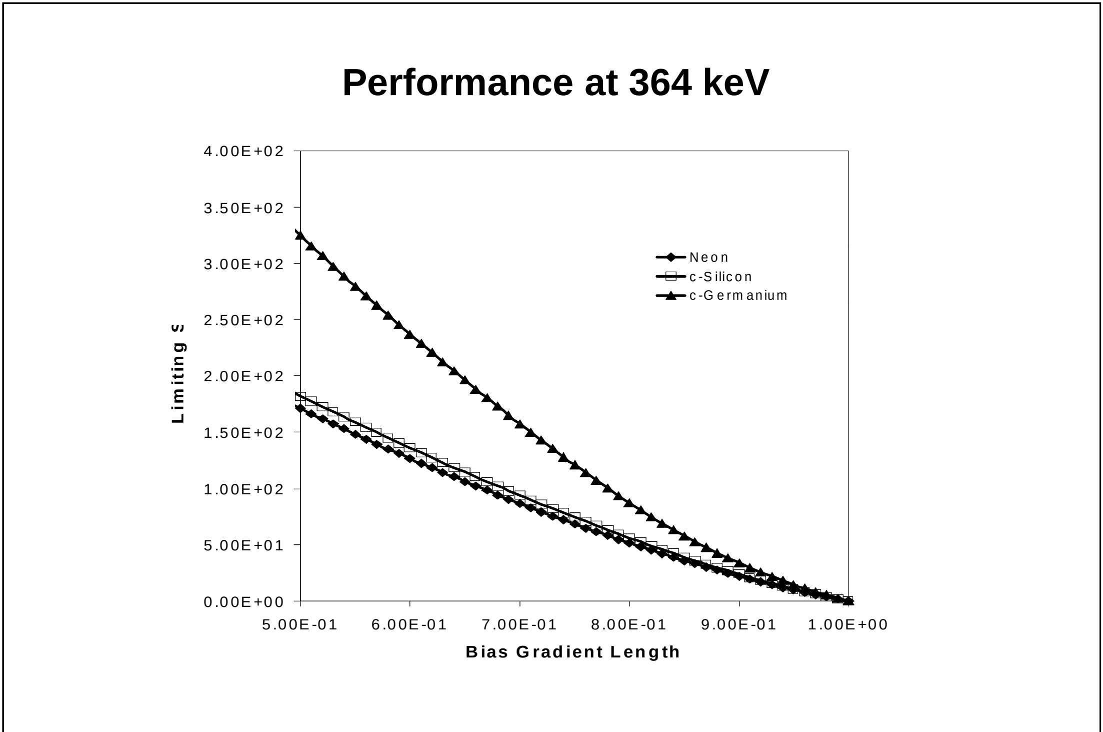
Object Parameterization

- *Planar objects, 64x64 pixel grid*
- *30 x 30 cm field-of-view*
- *Objects located 10 cm from collimator or scatter-ing detector*
- *Bound calculations focused on estimating cen- tral pixel intensity*

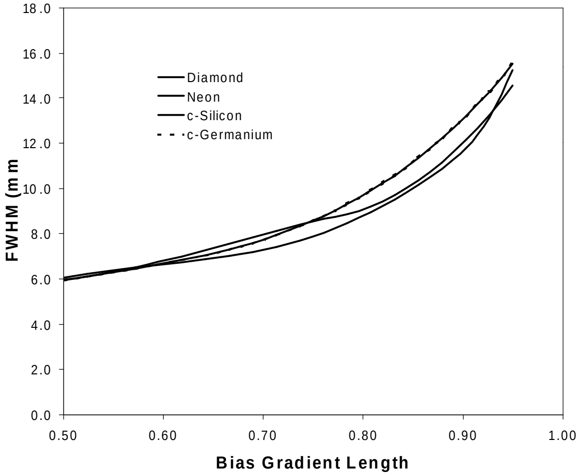
Bound Calculations

- ***Compared performance at 141 and 364 keV for several detector materials (diamond, c-silicon, c-germanium, neon).***
- ***Ignore potential energy resolution of each material (second detector can have good energy resolution).***
- ***Determined ability to estimate the intensity of the central pixel in a 7.5 cm diameter disk source 10 cm from detector.***
- ***Calculated using scattering angles from 45 through 90 degrees.***
- ***Calculations performed on 8 nodes of IBM SP2 computer. Approximately one week for each lower bound curve.***


Performance at 141 keV
Limiting SD
1.00E+03
9.00E+02
8.00E+02
7.00E+02
6.00E+02
5.00E+02
4.00E+02
3.00E+02
2.00E+02
1.00E+02
0.00E+00
5.00E-01
6.00E-01
7.00E-01
8.00E-01
9.00E-01
1.00E+00
Bias Gradient Length
Diamond
Neon
c-Silicon
c-Germanium


Performance at 364 keV
4.00E+02
3.50E+02
3.00E+02
2.50E+02
2.00E+02
1.50E+02
1.00E+02
5.00E+01
0.00E+00
Limiting ς
5.00E-01
6.00E-01
7.00E-01
8.00E-01
9.00E-01
1.00E+00
Bias Gradient Length
Neon
c-Silicon
c-Germanium

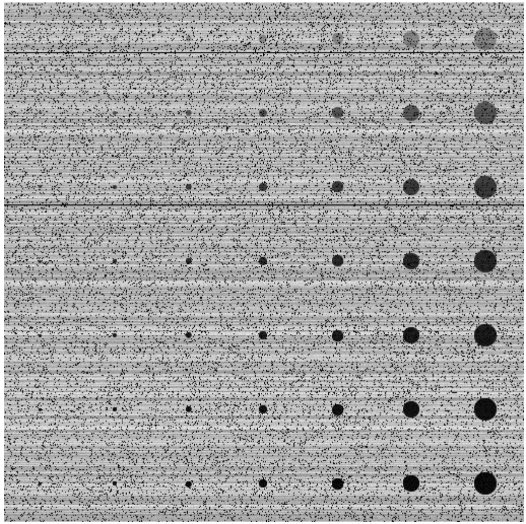
FWHM vs. Bias Gradient Length



Thinking Outside the Box

***Problem:** Pixellated detectors can suffer a variety of ills including significant gain variations among pixels as well as totally unresponsive pixels.*

Pictured at right is a simulated detector in which 10% of the pixels are unresponsive and two entire rows are missing.



Solution

Develop a statistical model describing the measurements, their noise, and potentially other degradations. Use a <u>penalized</u> least-squares method to estimate the incident image.

$$\nabla^{10} L(\theta, \mathbf{y}) = \left(\nabla \overline{\mathbf{y}}(\theta)\right)^{\mathrm{T}} \mathbf{K}_{\mathbf{y}}^{-1}\left[\mathbf{y} - \overline{\mathbf{y}}(\theta)\right] + \beta \mathbf{R}\theta$$

$$= \mathbf{G}\mathbf{K}_{\mathbf{y}}^{-1}\left[\mathbf{y} - \mathbf{G}\theta\right] + \beta \mathbf{R}\theta = 0$$

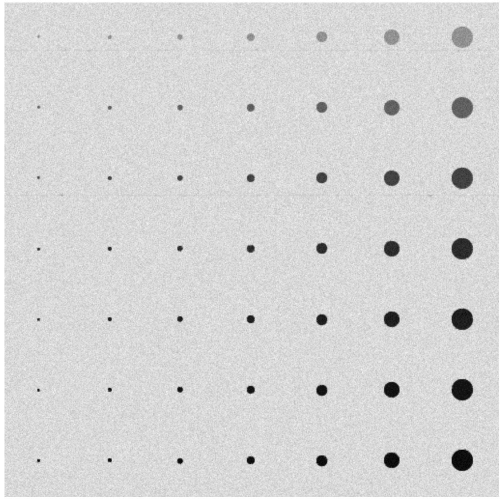
$$\mathbf{G} = diag(\text{Pixel Gains})$$

$$\mathbf{K}_{\mathbf{y}} = diag(\text{Pixel Noise})$$

This technique would not be very interesting if it weren't for the penalty matrix, which penalizes the differences between pixels to the North, Sourh, East, and West.

Corrected for Gain Variations and Dead Pixels

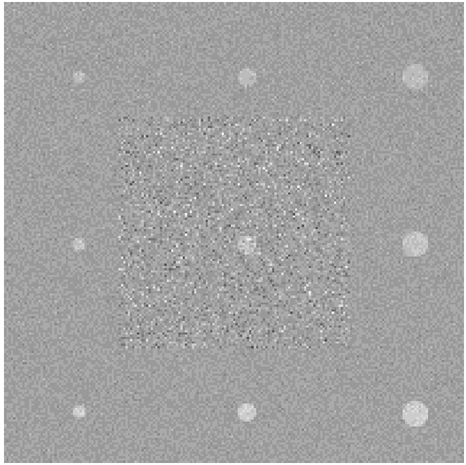
Applying the method to the previous image not only corrects for gain variations but also eliminates the effects of dead pixels.



Even More Problems

Regions of the detector may contain some pixels having significantly more noise than others.

In the central region, half the pixels have an additive (electronic) noise level 10x higher than others.



Corrected

This too, can be corrected by statistically modeling the measurements and using a correction method consistent with the model