WFS@LLL
(Wavefront sensing at low light level)

Harrison H. Barrett
University of Arizona

$\boxed{\text{OUTLINE}}$

- Review: Notation, bias, variance, MSE, FIM, CRB, ML, ...
- Statistical models
  Poisson models for photon statistics
  Gaussian models for read noise and dark current
  Mixed Poisson-Gaussian
- Application to the quad-cell Shack-Hartmann sensor
  Geometry
  What to estimate?
  Why not centroids?
  Form of the log-likelihood
  Expressions for the Fisher information matrix
- Numerical results
- Computational tricks (if time permits)

References

Barrett and Myers, FOUNDATIONS OF IMAGE SCIENCE
Chap. 13, Statistical Decision Theory

DETECTORS FOR SMALL-ANIMAL SPECT
II. Statistical Limitations and Estimation Methods
H. H. Barrett, Chapter in a forthcoming book (Kluwer, 2005)

The latter reference includes lot of discussion of lookup tables, data reduction, search methods and other computational tricks.

Notation and terminology

The $M \times 1$ vector g describes the random data.

The PDF on g is characterized by the $N \times 1$ parameter vector $\boldsymbol{\theta}$ and denoted pr$(\mathrm{g}|\boldsymbol{\theta})$. This PDF describes the *sampling distribution* of g for a given $\boldsymbol{\theta}$.

Once a data vector is measured, pr$(\mathrm{g}|\boldsymbol{\theta})$ can be regarded as a function of $\boldsymbol{\theta}$, sometimes called the *likelihood*:

$$L(\boldsymbol{\theta}|\mathrm{g}) = \text{pr}(\mathrm{g}|\boldsymbol{\theta}) = \text{pr}(\text{data}|\text{parameter})$$

Note that $L(\boldsymbol{\theta}|\mathrm{g})$ is *not* a PDF on $\boldsymbol{\theta}$.

An *estimate* of the parameter is denoted $\hat{\boldsymbol{\theta}}$; in most cases the estimate is a deterministic funcntion of the data, so we can also write it as $\hat{\boldsymbol{\theta}}(\mathrm{g})$. Since g is random (even for a given $\boldsymbol{\theta}$), so is $\hat{\boldsymbol{\theta}}(\mathrm{g})$.

Bias, variance, covariance and MSE

A real N-dimensional parameter vector $\boldsymbol{\theta}$ has an estimate $\hat{\boldsymbol{\theta}}$ with mean:

$$\overline{\hat{\boldsymbol{\theta}}} = \int d^M g \; \mathsf{pr}(\mathbf{g}|\boldsymbol{\theta}) \, \hat{\boldsymbol{\theta}}(\mathbf{g}) \, . \tag{13.282}$$

The bias $\mathbf{b}(\boldsymbol{\theta})$ is a vector quantity:

$$\mathbf{b}(\boldsymbol{\theta}) \equiv \overline{\hat{\boldsymbol{\theta}}} - \boldsymbol{\theta} \, . \tag{13.283}$$

The variance of the n^{th} element of the random vector $\hat{\boldsymbol{\theta}}$ is given by

$$\mathsf{Var}(\hat{\theta}_n) \equiv \left\langle \left[\hat{\theta}_n - \overline{\hat{\theta}}_n \right]^2 \right\rangle_{\mathbf{g}[\boldsymbol{\theta}} \, ,$$

and the nn' element of the covariance matrix is

$$\left[\mathbf{K}_{\hat{\boldsymbol{\theta}}} \right]_{nn'} = \left\langle \left[\hat{\theta}_n - \overline{\hat{\theta}}_n \right] \left[\hat{\theta}_{n'} - \overline{\hat{\theta}}_{n'} \right] \right\rangle$$

The mean-square error (MSE) is:

$$\mathsf{MSE} = \left\langle ||\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2 \right\rangle_{\mathbf{g}|\boldsymbol{\theta}} = \mathsf{tr} \left[\mathbf{K}_{\hat{\boldsymbol{\theta}}} \right] + \mathsf{tr} \left[\mathbf{b}\mathbf{b}^\dagger \right] \, .$$

Cramér-Rao bound for a scalar random variable

The variance of any unbiased estimate of a scalar must satisfy

$$\mathrm{Var}\{\hat{\theta}\} \geq \frac{1}{\left\langle \left[\frac{\partial}{\partial \theta} \ln \mathrm{pr}(\mathbf{g}|\theta) \right]^2 \right\rangle_{\mathbf{g}|\theta}}, \qquad (13.372)$$

where the denominator is called the *Fisher information*.

Similarly, for a biased estimator,

$$\mathrm{Var}\{\hat{\theta}\} \geq \frac{\left(\frac{db(\theta)}{d\theta} + 1 \right)^2}{\left\langle \left[\frac{\partial}{\partial \theta} \ln \mathrm{pr}(\mathbf{g}|\theta) \right]^2 \right\rangle_{\mathbf{g}|\theta}}. \qquad (13.377)$$

Fisher information in the vector case

For estimation of a scalar parameter, the Fisher information F is defined by:

$$F = \left\langle \left[\frac{\partial}{\partial \theta} \ln \mathrm{pr}(\mathbf{g}|\theta) \right]^2 \right\rangle_{\mathbf{g}|\theta} .$$

For a vector parameter with N components, the Fisher information $\mathbf{F}$ is an $N \times N$ Hermitian matrix with components:

$$F_{jk} = \left\langle \left[\frac{\partial}{\partial \theta_j} \ln \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) \right] \left[\frac{\partial}{\partial \theta_k} \ln \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) \right] \right\rangle_{\mathbf{g}|\boldsymbol{\theta}}$$

$$= \int_\infty d^M g \; \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) \left[\frac{1}{\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})} \frac{\partial}{\partial \theta_j} \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) \right] \left[\frac{1}{\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})} \frac{\partial}{\partial \theta_k} \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) \right] . \tag{13.361}$$

Note that the average is with respect to the same probability law as in the likelihood.

CR bound in the vector case

The nn component of the covariance matrix of any random vector is the variance of the n^{th} component. For any unbiased estimate,

$$[\mathbf{K}_{\hat{\boldsymbol{\theta}}}]_{nn} = \mathsf{Var}\{\hat{\theta}_n\} \geq \left[\mathbf{F}^{-1}\right]_{nn} . \tag{13.371}$$

Note that inversion of the Fisher information matrix is required to find the lower bound on the variance of a component of the estimate.

Other forms of the CR bound set limits on the covariance matrix of the estimate.

Terminology: An unbiased estimator that meets the CR bound is *efficient*.

Maximum-likelihood estimation

One general method of actually finding an estimate is *maximum-likelihood estimation*:

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \equiv \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \; \mathrm{pr}(\mathrm{g}|\boldsymbol{\theta})\,. \tag{13.348}$$

This procedure can be written equivalently as

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \; \ln[\mathrm{pr}(\mathrm{g}|\boldsymbol{\theta})]\,. \tag{13.349}$$

Note that we are *not* maximizing the probability of $\boldsymbol{\theta}$; we are choosing the value of $\boldsymbol{\theta}$ that maximizes the probability of occurrence of the g that we actually observed.

Why ML?

An ML estimate is:

- Efficient if an efficient estimate exists
- Asymptotically efficient (as you get more or better data)
- Asymptotically unbiased
- Asymptotically consistent
- Usually easy to compute, intuitive
- A way of rigorously enforcing agreement with the data
- A way of doing estimation with no prior information

Statistical limitations in wavefront sensing

- Photon noise
  Photons from the guide star
  Sky background
  Thermal radiation (in infrared imaging)

- Electronic noise
  Readout noise and dark current in detectors
  Amplifier noise

Note that the atmosphere is *not* a statistical limitation in wavefront sensing. Its characteristics are the parameters to be estimated.

Also, in ML estimation, the parameters themselves are regarded as unknown but not random. Thus MSE is a fair figure of merit but EMSE (ensemble-averaged MSE) is not.

$\boxed{\text{Photon noise}}$

POISSON $\Leftrightarrow$ INDEPENDENT

- Photons from the guide star are independent of each other
- Photons from the guide star are independent of sky background
- Photons from the guide star are independent of thermal radiation
- Photoelectrons in a detector are independent if the photons are
- Photoelectrons in one detector are independent of those in another

Punchline: For a given guide star, state of the atmosphere and scene being imaged, the set of photocounts in an array of detectors is a multivariate Poisson random vector.

Log-likelihood for Poisson data

A general multivariate Poisson data set has a probability (not PDF) given by

$$\Pr(\mathbf{g}) = \prod_{m=1}^{M} \exp(-\overline{g}_m) \frac{[\overline{g}_m]^{g_m}}{g_m!} \, .$$

Since this probability has only one parameter (the mean) for each m, the general Poisson likelihood is

$$\Pr(\mathbf{g}|\boldsymbol{\theta}) = \prod_{m=1}^{M} \exp[-\overline{g}_m(\boldsymbol{\theta})] \frac{[\overline{g}_m(\boldsymbol{\theta})]^{g_m}}{g_m!} \, ,$$

and the log-likelihood is

$$\ln \Pr(\mathbf{g}|\boldsymbol{\theta}) = \sum_{m=1}^{M} \{-\overline{g}_m(\boldsymbol{\theta}) + g_m \ln[\overline{g}_m(\boldsymbol{\theta})] - \ln g_m!\} \, .$$

Fisher information matrix for Poisson data

When the mean of an M-dimensional Poisson random vector is an arbitrary function of $\boldsymbol{\theta}$, it turns out (derivation on request) that the Fisher information has components:

$$F_{jk} = \sum_{m=1}^{M} \frac{1}{\overline{g}_m(\boldsymbol{\theta})} \frac{\partial \overline{g}_m(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \overline{g}_m(\boldsymbol{\theta})}{\partial \theta_k}$$

All you ever need to know to compute the likelihood or the Fisher information with Poisson data is $\overline{g}_m(\boldsymbol{\theta})$.

Electronic noise

- Electronic noise comes from lots of electrons

- A sum of N random variables tends to a Gaussian as $N \to \infty$ (central limit theorem). Here N is number of contributing electrons

- Noise in one detector independent of noise in another

- We often assume that all detectors are identical

- If there are several independent noise sources in each detector channel, overall variance $=$ sum of individual variances

Punchline: Electronic noise can be modeled as zero-mean i.i.d. Gaussian, and the overall variance can be assumed known.

Log-likelihood for i.i.d. Gaussian data

A general multivariate normal PDF has the form:

$$\text{pr}(\mathbf{g}) = \left[(2\pi)^M \det(\mathbf{K})\right]^{-1/2} \exp\left[-\tfrac{1}{2}(\mathbf{g}-\overline{\mathbf{g}})^t\,\mathbf{K}^{-1}(\mathbf{g}-\overline{\mathbf{g}})\right] , \qquad (8.185)$$

where $\overline{\mathbf{g}}$ is the mean vector and $\mathbf{K}$ is the covariance matrix of $\mathbf{g}$.

If the noise is i.i.d. with known variance σ^2 in each measurement, then $\mathbf{K} = \sigma^2\mathbf{I}$ and

$$\text{pr}(\mathbf{g}) = (2\pi\sigma^2)^{-M/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{m=1}^{M}(g_m - \overline{g}_m)^2\right] .$$

Since the variance is given, only the means $\overline{g}_m$ can depend on the parameters to be estimated, so the loglikelihood is given by

$$\ln \text{pr}(\mathbf{g}|\boldsymbol{\theta}) = constant - \frac{1}{2\sigma^2}\sum_{m=1}^{M}\left[g_m - \overline{g}_m(\boldsymbol{\theta})\right]^2 .$$

$\boxed{\text{Fisher information matrix for i.i.d. Gaussian data}}$

For i.i.d. Gaussian data, it turns out (derivation an easy exercise for the student) that the Fisher information has components:

$$F_{jk} = \frac{1}{\sigma^2} \sum_{m=1}^{M} \frac{\partial \overline{g}_m(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \overline{g}_m(\boldsymbol{\theta})}{\partial \theta_k}$$

As with Poisson data, all you ever need to know to compute the likelihood or the Fisher information is $\overline{g}_m(\boldsymbol{\theta})$.

Mixed Poisson and Gaussian noise

What happens when there is both electronic and Poisson noise?

Suppose the m^{th} detector receives k_m photoelectrons in exposure time T, responds to each with responsivity R [Volts/photon], and feeds the result into a readout channel with noise variance σ^2 [Volts^2]. The output of the electronics channel is denoted g_m, and its PDF is given by

$$\mathrm{pr}(g_m|\boldsymbol{\theta}) = \sum_{k_m=0}^{\infty} \mathrm{pr}(g_m|k_m)\, Pr(k_m|\boldsymbol{\theta})\,,$$

where $\mathrm{pr}(g_m|k_m)$ is the Gaussian PDF of the electronic signal for a fixed input and $Pr(k_m|\boldsymbol{\theta})$ is the Poisson probability (not PDF) for the photoelectrons. Note that only the latter depends on $\boldsymbol{\theta}$.

A closed form for $\mathrm{pr}(g_m|\boldsymbol{\theta})$ is not possible in general

Fisher information matrix for mixed Poisson and Gaussian noise

In the mixed case, it turns out (tricky derivation) that the components of the Fisher information are given approximately by

$$F_{jk} \approx \sum_{m=1}^{M} \frac{R^2}{\sigma^2 + R^2\,\overline{k}_m(\boldsymbol{\theta})} \frac{\partial \overline{k}_m(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \overline{k}_m(\boldsymbol{\theta})}{\partial \theta_k} ,$$

where $\overline{k}_m(\boldsymbol{\theta})$ is the mean number of photelectrons.

This expression is:

- Exact for i.i.d. Gaussian noise (lousy detectors)
- Exact for pure Poisson noise (perfect photon counters)
- An excellent approximation in the general case if $\overline{k}_m(\boldsymbol{\theta}) > 3$

Now all you need to know to compute the Fisher information matrix is $\overline{k}_m(\boldsymbol{\theta})$ (plus R and σ^2, of course).

Application of ML to wavefront sensing

Previous work:

R. G. Paxman, T. J. Schulz and J. R. Fienup, "Joint estimation of object and aberrations by using phase diversity" JOSA A, July 1992 [Have to check, but I think this is ML]

Mats G. Löfdahl and Alan L. Duncan, "Fast Phase Diversity Wavefront Sensing for Mirror Control" SPIE 3353 [Extended object, Gaussian model]

T.Y. Chew, R.M. Clare and R.G. Lane "A Cramer-Rao Bound analysis of the Shack-Hartmann and pyramid wavefront sensors" [Inappropriate likelihood model]

Scott A. Sallberg, Byron M. Welsh and Michael C. Roggemann "Maximum a posteriori estimation of wave-front slopes using a ShackHartmann wave-front sensor" JOSA A, June 1997 Poisson model, prior accounts for correlations of tilts across subapertures. Claims MAP is unbiased for centroid estimation]

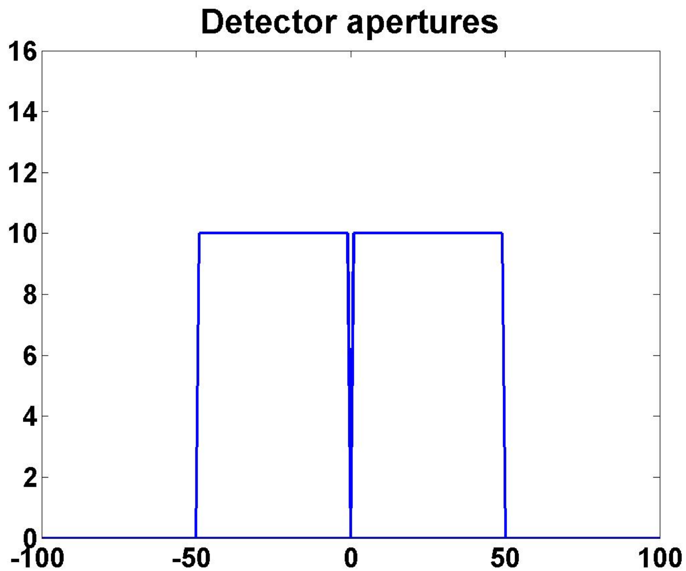
$\boxed{\text{Model problem: The quad-cell Shack-Hartmann}}$

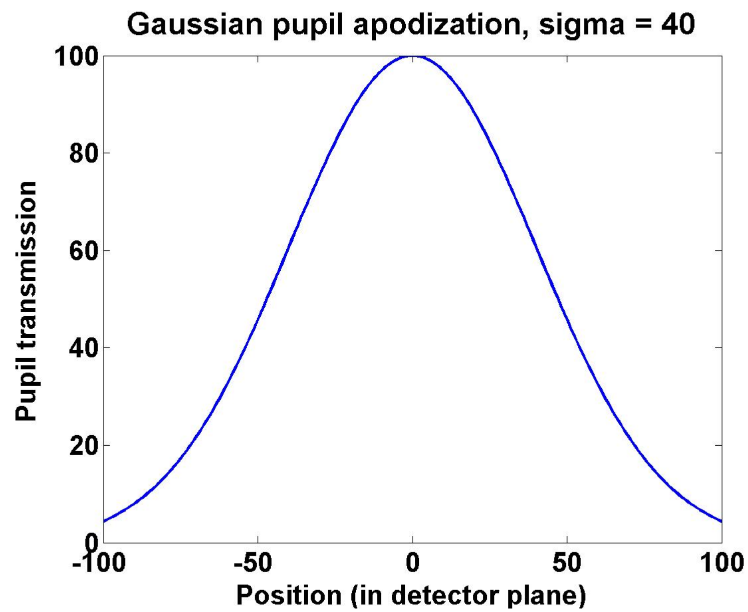
Assumptions:

- Pure tilt of the wavefront
- Square pupil, Gaussian or triangular apodization
- Spot blur caused by defocus
- Blur function given by scaled version of apodization
- Four identical square detectors

Detector and pupil functions

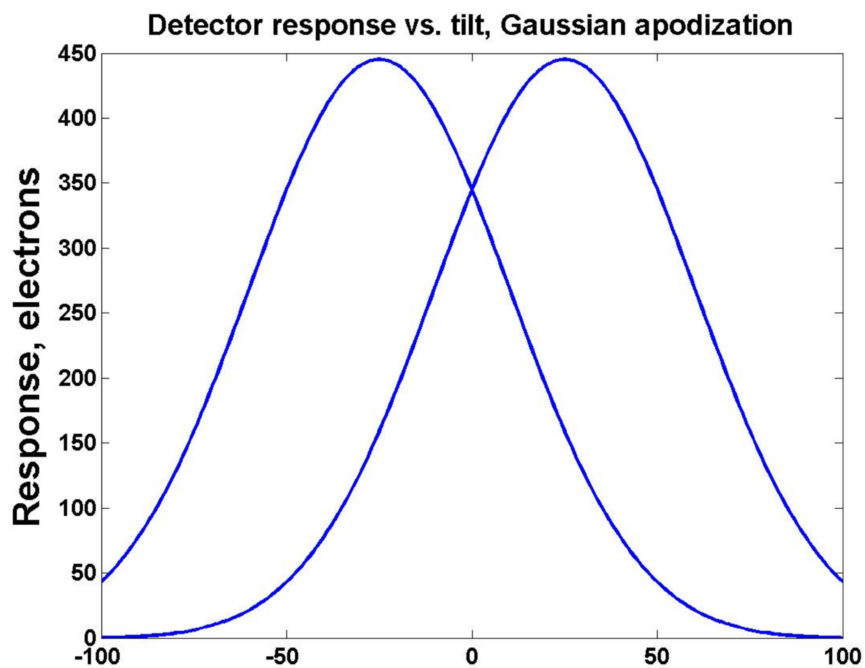
Gaussian apodization, square detectors



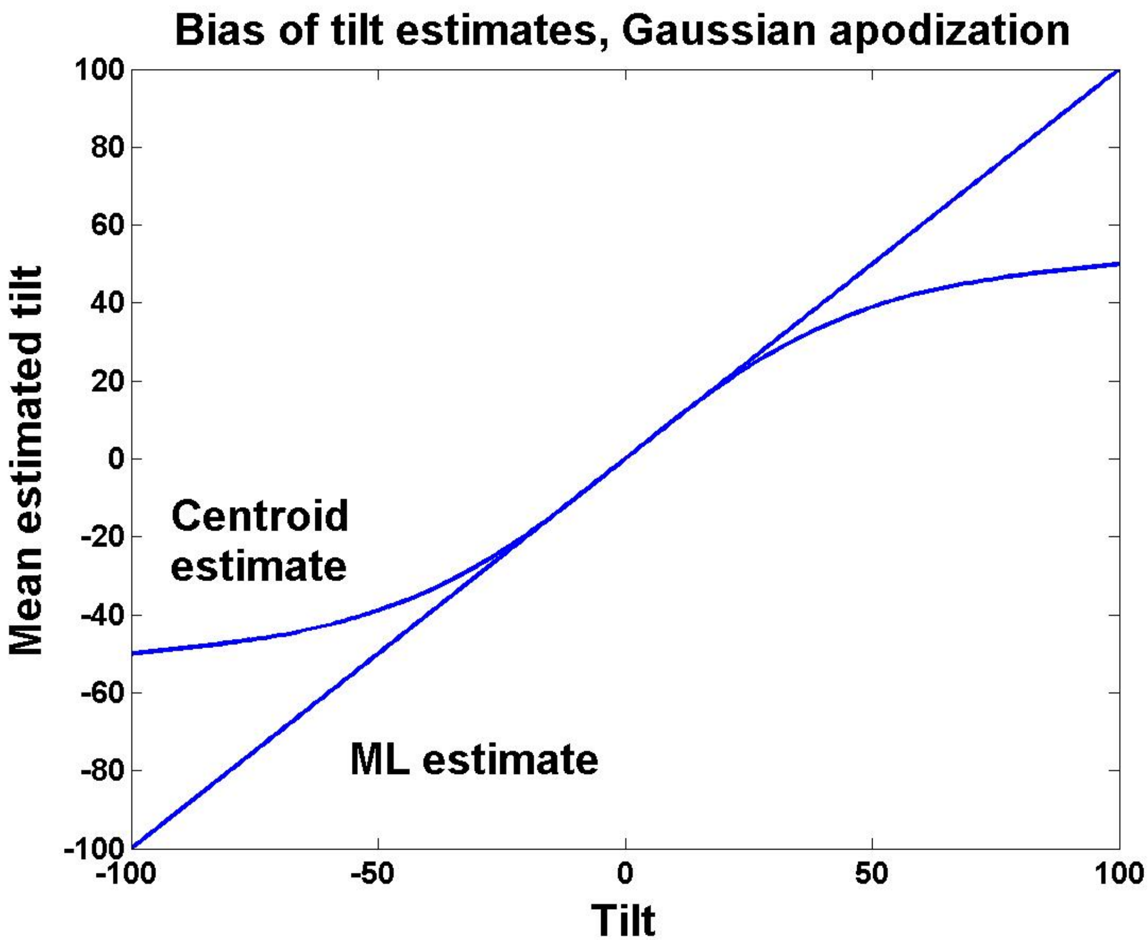


Detector response functions

Give or take some magnification factors, the mean detector response functions are convolutions of the pupil function with the detector apertures.



Why not centroids?



Fisher-information study of WFS design

What to vary:

- Apodizing function
- Amount of defocus
- Photon flux on pupil
- Variance of electronic noise

What to compute:

- Log-likelihood and FIM
- CR bound on tilt estimates
- CR bound averaged over tilt range
- Bias and variance of ML tilt estimates

Nuisance parameters

Even though we are interested only in the two components of tilt, other parameters, called *nuisance parameters*, may be needed to fully describe the data.

Examples:

- Brightness of guide star
- Sky background
- Residual wavefront curvature

What to do about nuisance parameters?

- Estimate them and throw away the result

- Measure them independently

- Assign some typical value

- Assume a prior and marginalize

- Ignore the problem

For this discussion, the only nuisance parameter we will consider is brightness of the guide star. We will either:

- Assume it is known independently, say from long exposure

or

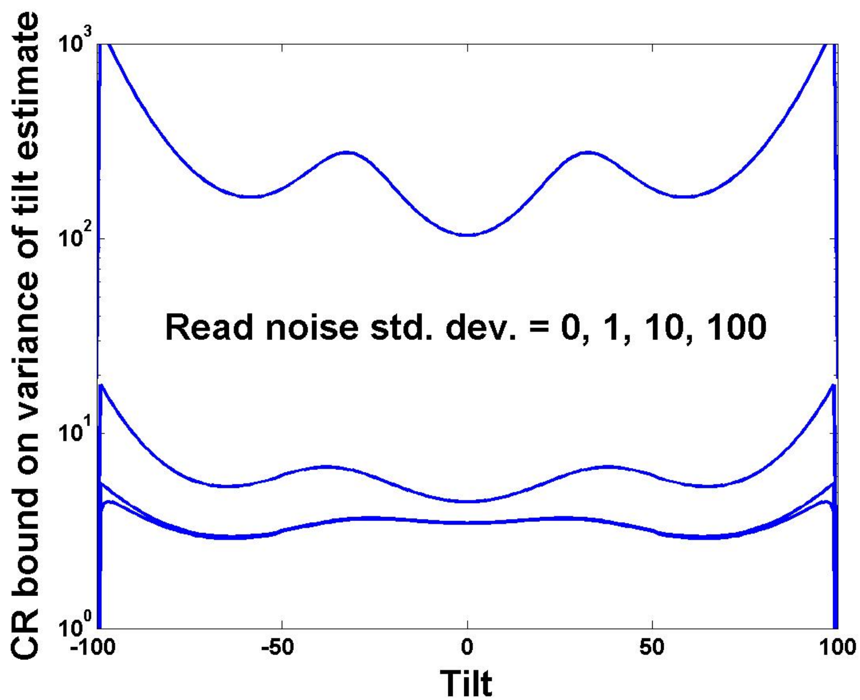
- Estimate it along with the tilt

In the latter case, the Fisher information matrix is 3×3 instead of 2×2

CR bounds

Assume that the brighness of the guide star is known *a priori*, so FIM is 2×2.

With our assumptions about the geometry, the x and y components of tilt enter symmetrically, and the FIM is diagonal. Inversion is trivial, and typical plots of CR bound vs. true tilt look like:



Lookup tables

For the quad cell, there are just four measurements. If each is made to B bits of accuracy, all possible measurements can be specified by $4B$ bits. If this number is small enough, it is feasible to precompute all possible ML estimates and store them in a lookup table.

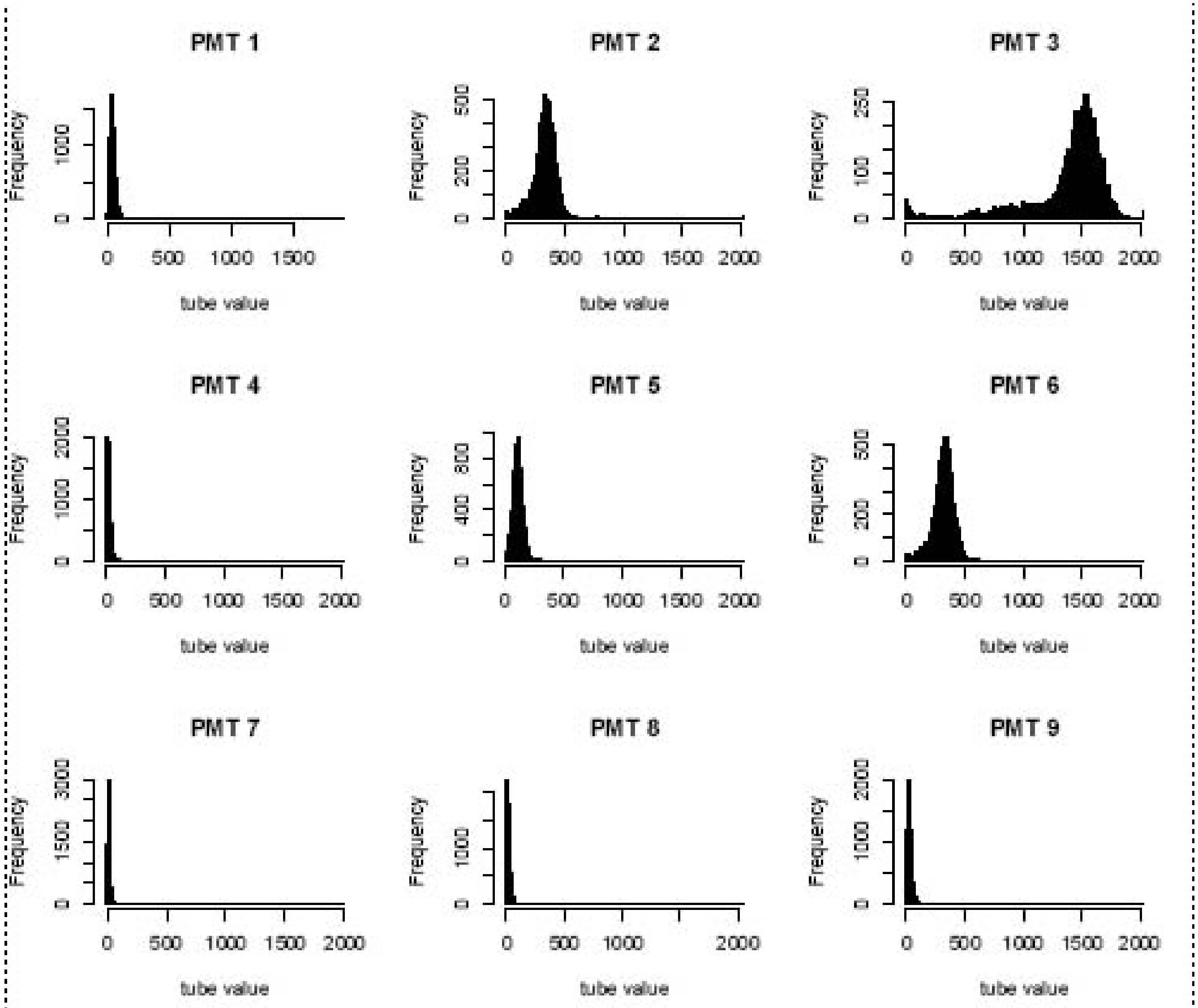
How big is the table?

Example 1: $B = 8$, $2^{4B} = 2^{32} \approx 4 \times 10^9$. To store estimates of x and y components of tilt at one byte each requires 8 GB of memory

Example 2: $B = 6$, $2^{4B} = 2^{24} \approx 16 \times 10^6 \Rightarrow 32$ MB of memory.

How many bits do we need?

Some measured histograms from a scintillation camera



The Anscombe transform

Let N be a Poisson random variable:

$$\mathsf{Var}(N) = \overline{N}$$

Now let

$$y = \sqrt{N}$$

It turns out that

$$\mathsf{Var}(y) \approx \frac{1}{4}$$

As a result, fewer bits are needed after the square-root transformation

Pictorial look at Anscombe

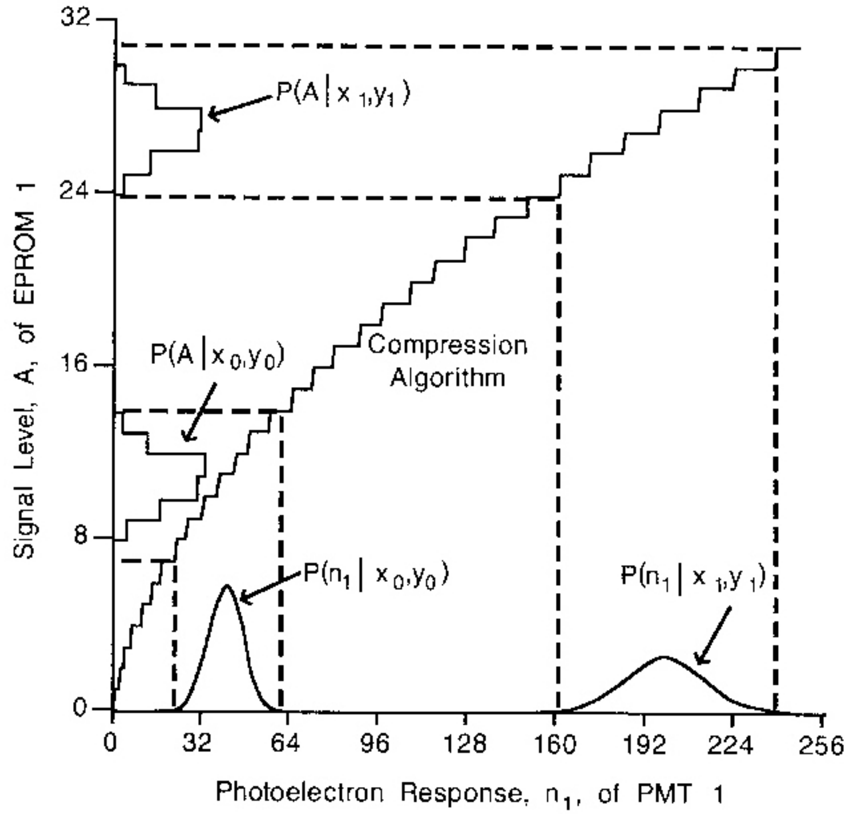


FIGURE 4

Transformation of $P(n_1|x,y)$ into $P(A/x,y)$. The Poisson probability distribution for n_1 is shown for two locations on the crystal, (x_0,y_0) and (x_1,y_1). These distributions are mapped into $P(A|x_0y_0)$ and $P(A|x_1y_1)$ by the binning and nonlinear compression algorithm. The variance of $P(n_1|xy)$, which is indicated by the width of the distribution in the figure, changes with (x,y). The variance of $P(A|xy)$ is nearly uniform with (x,y).

Still to do

- Verify experimentally that MLE approx. unbiased and efficient
- Carry out a full optimization of the quad-cell SHS
  - Optimal apodization
  - Optimal defocus
- Consider larger detector arrays
- Investigate the nuisance parameters more fully
- Consider joint tilt and curvature estimation
- Consider other kinds of WFS
- Implement the ideas in practice