Maximum-likelihood methods in wavefront sensing: stochastic models and likelihood functions

Harrison H. Barrett

College of Optical Sciences and Department of Radiology, University of Arizona, Tucson, Arizona 85724

Christopher Dainty

Department of Physics, National University of Ireland, Galway, Ireland

David Lara

Department of Physics, National University of Ireland, Galway, Ireland

Received November 22, 2005; revised May 2, 2006; accepted July 28, 2006;
posted September 5, 2006 (Doc. ID 66145); published January 10, 2007

Maximum-likelihood (ML) estimation in wavefront sensing requires careful attention to all noise sources and all factors that influence the sensor data. We present detailed probability density functions for the output of the image detector in a wavefront sensor, conditional not only on wavefront parameters but also on various nuisance parameters. Practical ways of dealing with nuisance parameters are described, and final expressions for likelihoods and Fisher information matrices are derived. The theory is illustrated by discussing Shack–Hartmann sensors, and computational requirements are discussed. Simulation results show that ML estimation can significantly increase the dynamic range of a Shack–Hartmann sensor with four detectors and that it can reduce the residual wavefront error when compared with traditional methods. © 2007 Optical Society of America

OCIS codes: 010.0010, 010.1080, 010.7350.

1. INTRODUCTION

Measurement of optical wavefronts has a long and storied history. Classical interferometry uses a reference beam to learn as much as possible about a wavefront, and phase-retrieval methods attempt to reconstruct a wavefront from one or more measurements of optical irradiance without a reference beam. In recent years, however, a distinctly different requirement has been imposed on systems for wavefront measurement: They have to respond to rapid changes in the wavefront and provide signals that can be used in adaptive systems that correct for wavefront distortions. Such adaptive systems are proving extremely valuable in many applications, including ground-based astronomy, retinal imaging in ophthalmology, and laser machining. In these applications there is no particular interest in the wavefront itself, but instead the goal is to sense a distorted wavefront, correct it, and thereby minimize its influence on the actual task of interest. Wavefront-measurement systems intended for use in adaptive optics (AO) are referred to as real-time wavefront sensors, or simply wavefront sensors for short.

Many different wavefront sensors have been developed for AO; for reviews, see Tyson[1] and Rousset.[2] The wavefront of interest is usually the pupil function of a telescope or other optical instrument, and the sensors differ in whether they attempt to characterize the wavefront over the entire pupil aperture at once or over selected regions called subapertures. All of the sensors, however, use a set of optical detectors in conjunction with optical elements intended to make the detector outputs sensitive to preselected characteristics of the wavefront. For example, the familiar Shack–Hartmann sensor attempts to measure two components of the wavefront tilt over a subaperture by observing the image of a star or other pointlike source in the back focal plane of a lenslet placed over the subaperture. Because of the lenslet, the image of the source is displaced laterally by an amount proportional to the tilt, and the displacement can be estimated by computing the centroid of the outputs of an array of detectors in the focal plane.

Other wavefront sensors attempt to measure other parameters, such as the local curvature of the wavefront at each subaperture[3] or the coefficients in an expansion of the wavefront in orthogonal basis functions over the whole aperture. Many clever techniques have been devised for choosing the configuration of optical elements and the photodetector array and for processing the outputs of the photodetectors to obtain measurements of the parameters of interest.

Most current real-time sensors can be described by the general block diagram shown in Fig. 1. The wavefront is assumed to be described by a set of P parameters $\{\theta_p,\ p=1,\ldots,P\}$, or equivalently by a $P\times 1$ parameter vector $\boldsymbol{\theta}$. Similarly, the raw data are described by a set of M output signals $\{g_m,\ m=1,\ldots,M\}$, or equivalently by an $M\times 1$ data vector $\mathbf{g}$. The photodetector signals are then prepro-

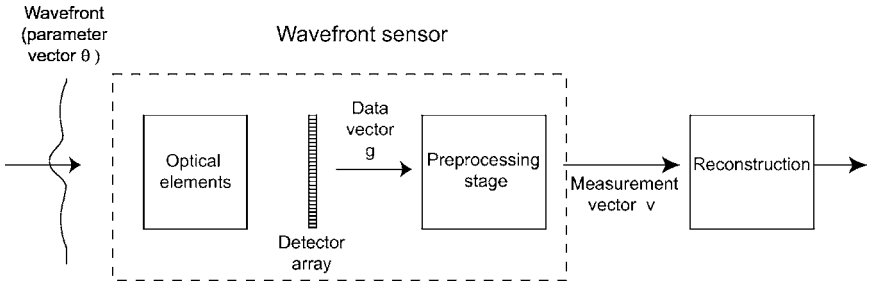

Fig. 1. Block diagram of a generic wavefront sensor and reconstructor.

cessed, usually by simple, noniterative formulas, to get a set of I derived quantities, $\{v_i,\ i=1,\ldots,I\}$ or an $I\times 1$ vector $\mathbf{v}$, that can be regarded as measurements of some properties of the wavefront, though not necessarily directly the components of $\boldsymbol{\theta}$. For example, in a Shack–Hartmann sensor for one subaperture, $I=2$ and the components of $\mathbf{v}$ are estimates of the tilts of the wavefront over the subaperture in the x and y directions. The preprocessing step in this case is computation of the centroid. Note that centroid computation, though fast and efficient, is a nonlinear operation on the data (because of the division by the sum of the signals).

No matter how the specific boxes in Fig. 1 are realized, it is usually assumed that there is a linear relation between the mean values of $\mathbf{v}$ and the actual wavefront parameters; this linear relation is expressed as

$$\bar{\mathbf{v}}=\mathbf{H}\boldsymbol{\theta} \quad \text{or} \quad \mathbf{v}=\mathbf{H}\boldsymbol{\theta}+\mathbf{n}, \tag{1.1}$$

where $\mathbf{H}$ is an $I\times P$ matrix, $\mathbf{n}$ is a zero-mean $I\times 1$ vector describing the noise in $\mathbf{v}$, and the overbar denotes an average over that measurement noise. Recovery of the unknown $\boldsymbol{\theta}$ from the output of the preprocessing stage is then treated as a matrix inversion or pseudoinversion implemented in a separate stage called a reconstructor. The output of the reconstructor can be the final estimates of $\boldsymbol{\theta}$ or correction signals to be applied to a control element (deformable mirror or spatial light modulator) in an AO system.

There are several difficulties with this general approach. An immediate concern is the linearity assumed in Eq. (1.1). Even in our example of a Shack–Hartmann sensor and centroid estimation, it is well known that the mean centroid is a nonlinear function of the tilts if the number of photodetectors is small. Moreover, if wavefront parameters other than tilt influence the data, then there is no chance that Shack–Hartmann tilt estimates will be linear functions of the additional parameters.

A more serious issue concerns the dimensionality reduction in going from the M-dimensional raw data $\mathbf{g}$ to the I-dimensional vector $\mathbf{v}$; as I is often much less than M, there could be a considerable information loss in this step. In the Shack–Hartmann example, we can expect wavefront curvature and other parameters to influence the data unless the lenslet diameter D_l is significantly smaller than the Fried parameter r_0. The usual choice, however, is to make D_l approximately equal to the mean r_0 at a particular observing site, and it is not clear in that case how much information is lost in centroid estimation.

A related problem is that parameters other than ones associated with the wavefront can influence the data. A simple example is the overall brightness of the guide star or other source, which is one additional scalar parameter. A more complex example is irradiance variations (e.g., scintillation) over the aperture being sensed, which would potentially require a large set of additional parameters. These extraneous parameters, called nuisance parameters, can have important effects on the data statistics.

In contrast to nuisance parameters, null functions are properties of the wavefront that might be of great interest but that do not influence the data. Since the matrix $\mathbf{H}$ in Eq. (1.1) has dimensions $I\times P$, with I often very small compared with P, there is a null space representing characteristics of $\boldsymbol{\theta}$ that cannot be recovered from knowledge of $\mathbf{v}$, even in the absence of noise.

Another area of difficulty is in describing the statistical properties of both $\mathbf{g}$ and $\mathbf{v}$. A centroid or other simple way of computing $\mathbf{v}$ from $\mathbf{g}$ takes no account of the noise properties of $\mathbf{g}$, and better performance might be obtainable if we used accurate models of the data statistics. Even if we do not use detailed statistical information in the preprocessing stage, it is still possible to compute the variances in the resulting components of $\mathbf{v}$ by simple propagation of errors[4] if we assume that the components of $\mathbf{g}$ are uncorrelated, but this assumption is not always justified.

Considerable work has been reported on optimal approaches to the reconstruction step, starting with the pioneering paper by Wallner.[5] This work starts with the assumption that the available data are noisy measurements of the wavefront tilts averaged over subapertures and that these measurements are unbiased and uncorrelated, both with each other and with the random wavefront itself. From this starting point, Wallner derives an optimal reconstructor that minimizes the mean-square wavefront error, accounting for unmeasured components by using Kolmogorov statistics as prior knowledge. His approach and subsequent related research thus optimize the reconstruction stage in Fig. 1, but they do not consider possible information loss in the preprocessing stage. As we shall demonstrate numerically in Section 6, that information loss can be considerable.

Moreover, the common assumption that the components of $\mathbf{v}$ are uncorrelated is almost never correct. Correlations are introduced by the preprocessing stage, and the statistics of $\mathbf{v}$ can be complicated, even when $\mathbf{g}$ is described by simple uncorrelated Gaussian or Poisson noise. At the least, any discussion of the statistics of the wavefront sensor output should give its mean (or bias), variance, and covariance matrix; a full multivariate probability density function would be desirable for rigorous design of the reconstruction stage.

Finally, there is a need for rigorous methods of evaluating wavefront sensors and comparing competing approaches. Most of the literature on this topic uses the Strehl ratio of the final AO system as the figure of merit, but it is difficult to discern the contribution of the wavefront sensor to this metric or to devise strategies for improving the sensor. Moreover, it is not clear how Strehl ratio itself relates to objective or task-based figures of merit[6–9] for the final system.

Likelihood theory offers a potential way of addressing all of these concerns. A likelihood is a comprehensive statistical description of a data set, showing how the data probability law depends on various parameters and various noise sources. This probability law can then be used to define a maximum-likelihood (ML) estimator, which has many desirable properties to be enumerated in Section 2. The likelihood is also required for Bayesian estimation methods, which augment the likelihood with prior knowledge of the parameters to be estimated.

From the likelihood it is possible to compute a Fisher information matrix (FIM), which describes the information content of a data set for the purpose of estimating the parameters that enter into the likelihood. It is well known that the FIM can be used to compute a fundamental lower bound, the Cramér–Rao bound (CRB), on the variance of the parameter estimates. It is less well known, but the FIM can also be used to find a good approximation to the covariance matrix of the ML estimates, and in this form it can be incorporated into objective theories of image quality.[8] In addition, likelihood theory provides a systematic way of discussing nuisance parameters.

Application of likelihood methods to wavefront sensing is not new, though their full potential has not yet been exploited. We can trace the beginnings of this line of research to three seminal 1974 papers by Bahaa Saleh,[10–12] in which he studied the statistical limitations in localizing a spot of light and derived ML estimators. Elbaum and Greenebaum[13] used similar methods for angular tracking, and Winick[14] derived a CRB for spot localization and used it to discuss system design. Various papers by Lane *et al.*[15–17] have applied ML methods and the CRB to wavefront sensors with the assumption that the positions of individual detected photons were available. Welsh *et al.*[18] used the CRB to compare the performance of Shack–Hartmann sensors and shearing interferometers. Löfdahl and Duncan[19] gave an ML treatment of the Shack–Hartmann sensor based on an additive Gaussian likelihood model, and they showed how to use the Shack–Hartmann for curvature estimation. Extension of ML methods to Bayesian MAP (maximum *a posteriori*) estimation is discussed by Sallberg *et al.*,[20] who used a Poisson likelihood and a prior on the correlation of wavefront slopes across subapertures in a Shack–Hartmann sensor.

An important paper by Cannon[21] considered ML estimation of global wavefront parameters from Shack–Hartmann data without the intermediary step of tilt estimation. His likelihood function took account of the polychromatic nature of the data, but it used an additive Gaussian model and did not consider photon noise.

Several papers[19,22–25] consider simultaneous ML or MAP estimation of a wavefront and an object from phase-diversity data without an explicit wavefront sensor; this problem does not fit into the general schema of Fig. 1, and it is not considered further in this paper.

Perhaps surprisingly, there is also some closely related work in a completely different area, namely gamma-ray detection with scintillation cameras in nuclear medicine. Gray and Macovski[26] suggested ML and MAP methods for localizing the spot of light produced by a single gamma ray in this application, and subsequent work at the University of Arizona and elsewhere[27–31] has refined the methodology and applied it to many practical gamma-ray imaging systems.

The objective of this paper is to develop rigorous likelihood models and FIMs for wavefront sensing under various noise assumptions and choices of parameters to estimate. In Section 2 we review some basic concepts in estimation theory, including the effect of null functions and nuisance parameters. In Section 3 we consider various stochastic models for the raw data in a WFS. These models are in the form of conditional probabilities or probability density functions (PDFs) on the photodetector outputs, conditioned on all parameters that influence those probabilities, but they are not yet likelihoods since we have not specified which of the parameters are to be estimated and how to handle those that will not be estimated. These topics are taken up in Section 4, where we consider various parametric descriptions of the wavefront and various choices of parameters to estimate. In Section 5 we combine the results from Sections 3 and 4 into practical likelihood functions and construct the corresponding FIMs. Section 6 applies these ideas specifically to a Shack–Hartmann sensor, and Section 7 discusses ways of finding ML estimates in a time compatible with astronomical adaptive optics.

Appendixes A and B provide some statistical details needed in the main text, and Appendixes C and D examine statistical issues particular to a Shack–Hartmann sensor.

2. BASIC CONCEPTS IN ESTIMATION THEORY

Random data are described by a probability law with one or more free parameters, and the goal of estimation is to obtain numerical values for the parameters from a given data set. Excellent general references on estimation theory include Melsa and Cohn,[32] Van Trees,[33] and Scharf.[34] An overview using a notation and approach similar to this paper is given by Barrett and Myers,[6] Chap. 13.

A. Notation and Terminology

Let $\mathbf{g}$ be an $M\times 1$ vector describing random data. The probability law on $\mathbf{g}$ is a PDF denoted $\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})$ for the case of continuous-valued data, and it is a probability denoted $\Pr(\mathbf{g}|\boldsymbol{\theta})$ for the case where the data can take on only discrete values. In both cases it is assumed that the probability law is characterized by a $P\times 1$ parameter vector $\boldsymbol{\theta}$. In the remainder of this section we shall consider continuous random variables, but the results are easily translated to discrete data.

The PDF describes the sampling distribution of the data, and we say that an individual sample of $\mathbf{g}$ is drawn

from $\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})$. Once a data vector is measured, however, $\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})$ can be regarded as a function of $\boldsymbol{\theta}$ called the likelihood of $\boldsymbol{\theta}$ for the given $\mathbf{g}$ and is denoted by

$$L(\boldsymbol{\theta}|\mathbf{g}) = \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}). \tag{2.1}$$

Note that $L(\boldsymbol{\theta}|\mathbf{g})$ 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 function of the data, so we can also write it as $\hat{\boldsymbol{\theta}}(\mathbf{g})$. Since $\mathbf{g}$ is random (even for a given $\boldsymbol{\theta}$), so is $\hat{\boldsymbol{\theta}}(\mathbf{g})$.

In wavefront sensing, we can choose either the raw photodetector output $\mathbf{g}$ or the derived quantities $\mathbf{v}$ as the data from which we wish to perform an estimation. In the latter case, the likelihood will be denoted $L(\boldsymbol{\theta}|\mathbf{v})$ or $\mathrm{pr}(\mathbf{v}|\boldsymbol{\theta})$, and an estimate will be denoted $\hat{\boldsymbol{\theta}}(\mathbf{v})$.

B. Performance Metrics

There are three distinct approaches to specifying the performance of an estimation procedure (or, indeed, any statistical inference task). There is the classical or frequentist method, which envisions repeated sampling of the data vector from its sampling distribution $\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})$ and bases its performance criteria on averages of the resulting estimates. In this view the parameter is unknown but not considered random. A Bayesian approach, on the other hand, considers the parameter being estimated to be random and assigns it a prior probability $\mathrm{pr}(\boldsymbol{\theta})$, though this probability may be regarded as a degree of belief rather than something that is necessarily verifiable by repeated experiments. By using $\mathrm{pr}(\boldsymbol{\theta})$ and $\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})$ in Bayes's rule, it is possible to assign a probability $\mathrm{pr}(\boldsymbol{\theta}|\mathbf{g})$, called the posterior to the value of $\boldsymbol{\theta}$ after the data vector is observed; all performance metrics are derived from the posterior.

The third approach to specification of estimation performance is to consider the use to which the estimate will be put. In an AO system, for example, we are not interested in the parameters of the wavefront but rather in the performance of the overall closed-loop system that uses the estimate. As noted in the introduction, a common way of specifying the overall performance in astronomical AO is in terms of Strehl ratio, but it is also possible to consider specific astronomical tasks such as detection of exoplanets and use a detectability measure as the final performance metric.[35] This approach is classical in the sense that it uses long-run averages, but they are averages related to the final task rather than to the estimates themselves.

In this paper we adopt the classical viewpoint. All probabilities and PDFs will be regarded as quantities that in principle can be verified by repeated sampling. Quantities like bias and variance of an estimator will thus have a frequentist (experimental) interpretation, but they will also serve as necessary inputs to a task-based assessment.

1. Bias, Variance, and Covariance of Estimates

In classical estimation theory, the accuracy of an estimate is specified in terms of its sampling distribution $\mathrm{pr}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})$, interpreted as the distribution of $\hat{\boldsymbol{\theta}}(\mathbf{g})$ that would be obtained by drawing repeated samples of $\mathbf{g}$ from $\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})$ and performing the estimation procedure on each. In terms of the sampling distribution, the mean of the $P\times 1$ vector of estimates is given by

$$\bar{\hat{\boldsymbol{\theta}}} = \int \mathrm{d}^P\hat{\boldsymbol{\theta}}\,\mathrm{pr}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})\hat{\boldsymbol{\theta}}. \tag{2.2}$$

If the estimation rule and the sampling distribution on $\mathbf{g}$ are known, we can also express the mean (expectation) of the estimate as

$$\bar{\hat{\boldsymbol{\theta}}} = \int \mathrm{d}^M g\,\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})\hat{\boldsymbol{\theta}}(\mathbf{g}) \equiv \left\langle \hat{\boldsymbol{\theta}}(\mathbf{g})\right\rangle_{\mathbf{g}|\boldsymbol{\theta}}. \tag{2.3}$$

We shall use the overbar and the angle brackets interchangeably to denote means; the latter has the advantage that the subscript can show explicitly which PDF is implied in the averaging process.

The *bias* in an estimate specifies its average deviation from the true value of the parameter. For a vector parameter, the bias is a vector given by

$$\begin{aligned}\mathbf{b}(\boldsymbol{\theta}) &\equiv \bar{\hat{\boldsymbol{\theta}}} - \boldsymbol{\theta}\\ &\equiv \int_\infty \mathrm{d}^M g\,[\hat{\boldsymbol{\theta}}(\mathbf{g}) - \boldsymbol{\theta}]\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})\\ &= \int_\infty \mathrm{d}^P\theta[\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}]\mathrm{pr}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}).\end{aligned} \tag{2.4}$$

A parameter is said to be be estimable or identifiable with respect to some data set if there exists an unbiased estimator of it for all true values of the parameter.

If we denote the mean of the pth element of the random vector $\hat{\boldsymbol{\theta}}$ by $\langle\hat{\boldsymbol{\theta}}_p\rangle$, the variance of the pth element is given by

$$\begin{aligned}\mathrm{Var}(\hat{\boldsymbol{\theta}}_p) &\equiv \left\langle\left[\hat{\boldsymbol{\theta}}_p - \left\langle\hat{\boldsymbol{\theta}}_p\right\rangle\right]\left[\hat{\boldsymbol{\theta}}_p - \left\langle\hat{\boldsymbol{\theta}}_p\right\rangle\right]^*\right\rangle_{\mathbf{g}|\boldsymbol{\theta}}\\ &= \int_\infty \mathrm{d}^M g|\hat{\boldsymbol{\theta}}_p(\mathbf{g}) - \left\langle\hat{\boldsymbol{\theta}}_p(\mathbf{g})\right\rangle|^2\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})\\ &= \int_\infty \mathrm{d}^P\theta\,|\hat{\boldsymbol{\theta}}_p - \left\langle\hat{\boldsymbol{\theta}}_p\right\rangle|^2\mathrm{pr}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta}),\end{aligned} \tag{2.5}$$

and the full covariance matrix is given by

$$[\mathbf{K}_{\hat{\boldsymbol{\theta}}}]_{pp'} = \left\langle\left[\hat{\boldsymbol{\theta}}_p - \left\langle\hat{\boldsymbol{\theta}}_p\right\rangle\right]\left[\hat{\boldsymbol{\theta}}_{p'} - \left\langle\hat{\boldsymbol{\theta}}_{p'}\right\rangle\right]^*\right\rangle_{\mathbf{g}|\boldsymbol{\theta}}$$

or

$$\mathbf{K}_{\hat{\boldsymbol{\theta}}} = \left\langle(\hat{\boldsymbol{\theta}} - \bar{\hat{\boldsymbol{\theta}}})(\hat{\boldsymbol{\theta}} - \bar{\hat{\boldsymbol{\theta}}})^\dagger\right\rangle_{\mathbf{g}|\boldsymbol{\theta}}, \tag{2.6}$$

where the dagger denotes adjoint (conjugate transpose), or simply transpose for real vectors and matrices.

2. Mean-Square error

The mean-square error (MSE) is a way of specifying the overall error, including bias and variance, in a single scalar quantity; it is defined by

$$\mathrm{MSE} = \left\langle \|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \right\rangle_{\mathbf{g}|\boldsymbol{\theta}} = \int_{\infty} \mathrm{d}^M g \left\| \hat{\boldsymbol{\theta}}(\mathbf{g}) - \boldsymbol{\theta} \right\|^2 \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) = \mathrm{tr}[\mathbf{K}_{\hat{\boldsymbol{\theta}}}] + \mathrm{tr}[\mathbf{b}\mathbf{b}^{\dagger}], \tag{2.7}$$

where tr(·) denotes the trace. Note that the MSE measures the squared deviation from the true value of the parameter, while the variances relate to deviations from the mean of the estimate.

In general, bias, variance, and MSE will all depend on the true value of the parameter. If a realistic sampling distribution of the parameter is known, it can be used to average the MSE, forming a quantity called the ensemble MSE, defined by

$$\mathrm{EMSE} = \left\langle \left\langle \|\hat{\boldsymbol{\theta}}(\mathbf{g}) - \boldsymbol{\theta}\|^2 \right\rangle_{\mathbf{g}|\boldsymbol{\theta}} \right\rangle_{\boldsymbol{\theta}}. \tag{2.8}$$

The EMSE can often be estimated by Monte Carlo sampling even when we do not have enough detail about the prior to use it in Bayesian estimation.

3. Cost and Risk

A general approach to estimation is to define a cost function $C(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})$ and to define the risk R as an average cost, $R = \langle C(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}) \rangle$. Depending on the statistical philosophy being adopted, the angle brackets here can have one of three distinct meanings. In a purely frequentist approach, the brackets imply averaging over $\mathbf{g}$ for a given $\boldsymbol{\theta}$, so the risk is a function of $\boldsymbol{\theta}$. In a purely Bayesian view, the average is over $\boldsymbol{\theta}$ for a given $\mathbf{g}$, so the risk is a function of the particular data set $\mathbf{g}$ and no other data set is ever considered. A pragmatic view is to average over both $\mathbf{g}$ given $\boldsymbol{\theta}$ and then over $\boldsymbol{\theta}$, so that the risk is a pure number. The EMSE in Eq. (2.8) is an example of risk defined this way for a quadratic cost function.

No matter what cost function and definition of risk are used, a nuisance parameter can be defined as one that does not appear in the cost function.

C. Nuisance Parameters and Null Functions

The performance metrics discussed above must be interpreted carefully when the measurement system has null functions or when there are nuisance parameters in the problem.

Null functions do not influence the data and in principle cannot be determined from the data. An example in the context of wavefront sensing is the piston component of the wave over a lenslet in a Shack–Hartmann sensor. We need to know this component to reconstruct the wavefront, but the sensor is not responsive to it. A second example is the so-called waffle effect, which arises when the deformable mirror in an AO system has modes that the wavefront sensor cannot detect; the resulting corrected wavefront then has a corrugated or waffled appearance.

Nuisance parameters do influence the data but are not of interest to the estimation problem, perhaps because they do not influence performance of the real task of interest. An example in astronomical applications is the brightness of the guide star. Like all nuisance parameters, the brightness of the guide star influences the bias and/or variance of the estimates of the parameters of interest, but the value of the brightness itself is irrelevant to further application of the output of the WFS. If there is atmospheric scintillation or if the guide star is laser-induced and hence noisy, however, fluctuations in the brightness can be a serious nuisance.

In a sense it is trivial to deal with null functions. Since they do not affect the data and cannot be estimated from the data, we can just omit them from the likelihood function and the FIM. On the other hand, if we do try to estimate them, for example by trying to solve Eq. (1.1) for the case $P > I$, then the FIM is singular[36] and the CRB is infinite. Stated differently, $\boldsymbol{\theta}$ is not estimable. This difficulty often goes unrecognized in the wavefront-sensing literature and in other areas of inverse problems.

In contrast to null functions, it is never correct to omit nuisance parameters from the likelihood, though in fact it is often done. A correct statistical description of the data has the form $\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})$, where the vector $\boldsymbol{\theta}$ contains all of the parameters that influence the data, not just those we might want to estimate.

Methods of dealing with nuisance parameters are summarized in Barrett and Myers.[6] If we write

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{pmatrix},$$

where $\boldsymbol{\alpha}$ contains the parameters of interest and $\boldsymbol{\beta}$ contains the nuisance parameters, we can

(1) Ignore the problem and assume a form for $\mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha})$.

(2) Replace $\boldsymbol{\beta}$ with some typical value $\boldsymbol{\beta}_0$ and assume that $\mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha},\boldsymbol{\beta}) \approx \mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha},\boldsymbol{\beta}_0)$.

(3) Estimate $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ simultaneously from $\mathbf{g}$ and discard the estimate of $\boldsymbol{\beta}$.

(4) Estimate $\boldsymbol{\beta}$ from some auxiliary data set and use it as in option (2).

(5) Assume (or measure) some prior $\mathrm{pr}(\boldsymbol{\beta})$ and marginalize over $\boldsymbol{\beta}$.

It is shown by Barrett and Myers[6] (Sec. 13.3.8) that option (5) is optimal in terms of minimizing a particular cost function (the one that leads to MAP estimation), provided that the cost is independent of the nuisance parameter. It is assumed there, however, that $\mathrm{pr}(\boldsymbol{\beta})$ is a meaningful sampling prior, not something based on belief or chosen for mathematical convenience. For a good discussion of marginalization from a Bayesian perspective, see Berger.[37]

These five approaches to dealing with nuisance parameters will be discussed further in the context of wavefront sensing in Section 5.

D. Fisher Information and Cramér-Rao Bounds

For a vector parameter with P real components, the FIM, denoted $\mathbf{F}$, is a $P \times P$ symmetric matrix with components given by

$$F_{jk} = \left\langle \left[\frac{\partial}{\partial \theta_j} \ln \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) \right] \left[\frac{\partial}{\partial \theta_k} \ln \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) \right] \right\rangle_{\mathbf{g}|\boldsymbol{\theta}}$$
$$= \int_{\infty} \mathrm{d}^M g \, \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) \left[\frac{1}{\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})} \frac{\partial}{\partial \theta_j} \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) \right]$$
$$\times \left[\frac{1}{\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})} \frac{\partial}{\partial \theta_k} \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) \right]. \quad (2.9)$$

Note that the FIM is fully determined by the likelihood function; it is the covariance matrix of the gradient of the logarithm of the likelihood, and the average itself is with respect to the likelihood function. In general the FIM will depend on the true parameter $\boldsymbol{\theta}$.

An important use of the FIM is to determine the lower CRB on the variance of the estimate. It is shown in any standard text[32,33] that the variance of any unbiased estimate must satisfy

$$[\mathbf{K}_{\hat{\boldsymbol{\theta}}}]_{nn} = \mathrm{Var}\{\hat{\boldsymbol{\theta}}_n\} \geq [\mathbf{F}^{-1}]_{nn}. \quad (2.10)$$

Note that inversion of the Fisher information is required to find the lower bound on the variance of a component of the estimate. An unbiased estimator that achieves the bound of inequality (2.10) is called "efficient."

Inequality (2.10) is a special case of a more general relation, which can be stated with the help of a notational convention known as Loewner ordering (see Barrett and Myers,[6] Appendix A). If we have two $P \times P$ positive-definite matrices $\mathbf{A}$ and $\mathbf{B}$, the statement $\mathbf{A} \geq \mathbf{B}$ does not hold on an element-by-element basis. Rather, it means that $\mathbf{A} - \mathbf{B}$ is positive-semidefinite, or equivalently that $\mathbf{x}^\dagger \mathbf{A} \mathbf{x} \geq \mathbf{x}^\dagger \mathbf{B} \mathbf{x}$ for all $\mathbf{x}$.

With this convention, it can be shown that the covariance matrix for any unbiased estimator must satisfy

$$\mathbf{K}_{\hat{\boldsymbol{\theta}}} \geq \mathbf{F}^{-1}. \quad (2.11)$$

The corresponding relation for a biased estimator is

$$\mathbf{K}_{\hat{\boldsymbol{\theta}}} \geq (\nabla_{\boldsymbol{\theta}} \mathbf{b} + \mathbf{I}) \mathbf{F}^{-1} (\nabla_{\boldsymbol{\theta}} \mathbf{b} + \mathbf{I})^t, \quad (2.12)$$

where $\mathbf{I}$ is the $P \times P$ unit matrix. Thus the bias of an estimator alters the lower bound on the variance by an amount that depends on the bias gradient. Note that bias can decrease the variance if the bias gradient is negative.

E. Maximum-Likelihood Estimation

So far we have not talked about ways of actually finding an estimate. One general method is ML estimation, defined by

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \equiv \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \, \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}), \quad (2.13)$$

where the argmax operator returns the $\boldsymbol{\theta}$ argument at which $\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})$ is maximized. Since the logarithm is a monotonic function of its argument, Eq. (2.13) can also be written as

$$\hat{\boldsymbol{\theta}}_{\mathrm{ML}} \equiv \underset{\boldsymbol{\theta}}{\mathrm{argmax}} \ln [\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})]. \quad (2.14)$$

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 $\mathbf{g}$ that we actually observed.

ML estimates have many desirable properties.[6,38] First, they are efficient if an efficient estimate exists for a particular problem. And even when no efficient estimator exists, the ML estimate is asymptotically efficient and asymptotically unbiased in a sense to be explained in the next paragraph. Moreover, the PDF on ML estimates, $\mathrm{pr}(\hat{\boldsymbol{\theta}}|\boldsymbol{\theta})$, is asymptotically a multivariate normal with the covariance matrix given by taking the equality sign in expression (2.11).

The asymptotic properties listed above are usually stated by assuming that N independent samples of $\mathbf{g}$ are drawn from the same $\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})$ and then letting $N \to \infty$; but in fact they hold also when one gets better data, for example by collecting more photons if the primary noise is Poisson or by letting the variance go to zero for Gaussian noise. With better data, therefore, the ML estimate approaches an efficient estimate, and its PDF approaches a fully specified multivariate normal law.

Another useful property of ML estimation arises when you want to estimate some function of the $\boldsymbol{\theta}$ that appears in the likelihood, rather than $\boldsymbol{\theta}$ itself. If we let $\mathbf{a}(\boldsymbol{\theta})$ be a prescribed one-to-one vector-valued function, then under mild conditions it can be shown that[34]

$$\hat{\mathbf{a}}_{ML} = \mathbf{a}(\hat{\boldsymbol{\theta}}_{ML}). \quad (2.15)$$

This property is referred to as the invariance of ML estimates.

3. STOCHASTIC DATA MODELS

In this section we present various probability laws for the raw data $\mathbf{g}$ (the output of the photodetector array in Fig. 1), and we briefly consider models for the derived measurements $\mathbf{v}$. The probability laws will depend on some set of parameters $\boldsymbol{\theta}$, so we shall give expressions for the conditional probability laws, $\mathrm{pr}(\mathrm{data}|\boldsymbol{\theta})$, along with the corresponding FIM that would be relevant if we wanted to estimate all components of $\boldsymbol{\theta}$. In practical applications such as wavefront sensing, however, we may not want (or be able) to estimate all components of $\boldsymbol{\theta}$. In Section 4 we shall look more closely at what we can and should estimate, and in Section 5 the probability laws presented in this section will be converted to practical likelihoods and FIMs.

A. Pure Poisson Statistics

If we consider an array of ideal photon-counting detectors and a radiation source that satisfies the conditions for Poisson statistics (see Barrett and Myers[6] for an extensive discussion), then g_m is the observed number of photocounts (photoelectric interactions) in the mth detector element. Similarly, dark current is frequently modeled as Poisson.

Since Poisson events are inherently independent and the Poisson probability is determined fully by its mean, the multivariate conditional probability on the data (the likelihood for estimation of $\boldsymbol{\theta}$) is given by

$$\Pr(\mathbf{g}|\boldsymbol{\theta}) = \prod_{m=1}^{M} \exp[-\bar{g}_m(\boldsymbol{\theta})] \frac{[\bar{g}_m(\boldsymbol{\theta})]^{g_m}}{g_m!}, \tag{3.1}$$

and its logarithm is

$$\ln \Pr(\mathbf{g}|\boldsymbol{\theta}) = \sum_{m=1}^{M} \{-\bar{g}_m(\boldsymbol{\theta}) + g_m \ln[\bar{g}_m(\boldsymbol{\theta})] - \ln(g_m!)\}. \tag{3.2}$$

If the vector $\boldsymbol{\theta}$ includes all parameters that influence the data, and all of these parameters are to be estimated, then Eq. (3.2) can be interpreted as a log-likelihood. The FIM in that case is readily derived from its definition [Eq. (2.9)].

The derivative of the log-likelihood with respect to a component of $\boldsymbol{\theta}$ is

$$\frac{\partial}{\partial \theta_j} \ln \Pr(\mathbf{g}|\boldsymbol{\theta}) = \sum_{m=1}^{M} \left\{ -1 + \frac{g_m}{\bar{g}_m(\boldsymbol{\theta})} \right\} \frac{\partial \bar{g}_m(\boldsymbol{\theta})}{\partial \theta_j}. \tag{3.3}$$

Poisson random variables are uncorrelated and have a variance equal to their mean,

$$\left\langle [g_m - \bar{g}_m(\boldsymbol{\theta})][g_{m'} - \bar{g}_{m'}(\boldsymbol{\theta})] \right\rangle = \bar{g}_m(\boldsymbol{\theta}) \delta_{mm'}, \tag{3.4}$$

so it follows from Eq. (2.9) and a little algebra that

$$F_{jk} = \sum_{m=1}^{M} \frac{1}{\bar{g}_m(\boldsymbol{\theta})} \frac{\partial \bar{g}_m(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \bar{g}_m(\boldsymbol{\theta})}{\partial \theta_k}. \tag{3.5}$$

To reiterate, these expressions for likelihood and FIM hold rigorously only if $\boldsymbol{\theta}$ includes all parameters that can influence the data (including, for example, the brightness of the guide star).

An example of the pure Poisson model occurs in the work of Winick,[14] who considered Poisson noise arising from a light spot projected onto a CCD detector and also from a dark current in the detector. The parameter vector $\boldsymbol{\theta}$ in his case consisted of just the x and y coordinates of the spot.

B. List-Mode Data

One interesting special case of Eq. (3.2) that has been considered in the literature on wavefront sensing[15–17] is the limit of very small detector elements. In that case, no element will detect more than one photon and the array will provide the coordinates of every detected photon. If K photons are detected, the data set, denoted $\mathbf{G}$ to distinguish it from the usual binned data, is a set of $K+1$ quantities, namely each 2D position vector $\mathbf{r}_k = (x_k, y_k)$ as well as K itself. This way of expressing information about a collection of photons is known in the nuclear-medicine literature as *list mode*; the coordinates and other parameters (e.g., time of arrival, photon energy if it can be measured) are stored in a list. List-mode likelihood and image reconstruction from list-mode data have been well studied in the medical literature.[39,40]

The likelihood for a photon list can be expressed as

$$\mathrm{pr}(\mathbf{G}|\boldsymbol{\theta}) = \mathrm{pr}(\{\mathbf{r}_k\}, K|\boldsymbol{\theta}) = \mathrm{pr}(\{\mathbf{r}_k\}|K, \boldsymbol{\theta}) \Pr(K|\boldsymbol{\theta}), \tag{3.6}$$

where $\mathrm{pr}(\{\mathbf{r}_k\}, K|\boldsymbol{\theta})$ is a multivariate PDF on the photon positions $\mathbf{r}_k$ but a probability on the discrete random variable K. Under the same assumptions that lead to the independent Poisson form in Eq. (3.1), the photons are independent, and we can write

$$\mathrm{pr}(\mathbf{G}|\boldsymbol{\theta}) = \Pr(K|\boldsymbol{\theta}) \prod_{k=1}^{K} \mathrm{pr}(\mathbf{r}_k|\boldsymbol{\theta}), \tag{3.7}$$

where $\mathrm{pr}(\mathbf{r}_k|\boldsymbol{\theta})$ is the PDF for the location of the kth photon; since the photons are indistinguishable, this PDF must be the same for all k. In fact, it is known from the theory of Poisson random processes[6] that

$$\mathrm{pr}(\mathbf{r}_k|\boldsymbol{\theta}) = \frac{b(\mathbf{r}_k; \boldsymbol{\theta})}{\int_{det} \mathrm{d}^2 r\, b(\mathbf{r}; \boldsymbol{\theta})}, \tag{3.8}$$

where $b(\mathbf{r}; \boldsymbol{\theta})$ is the photon fluence (the mean number of photons per unit area for parameter $\boldsymbol{\theta}$), and the integral is over the area of the detector array.

Since K is a Poisson random variable, the likelihood for the list is given by

$$\begin{aligned} \mathrm{pr}(\mathbf{G}|\boldsymbol{\theta}) &= \exp[-\bar{K}(\boldsymbol{\theta})] \frac{[\bar{K}(\boldsymbol{\theta})]^K}{K!} \prod_{k=1}^{K} \frac{b(\mathbf{r}_k; \boldsymbol{\theta})}{\int_{det} \mathrm{d}^2 r\, b(\mathbf{r}; \boldsymbol{\theta})} \\ &= \frac{\exp[-\bar{K}(\boldsymbol{\theta})]}{K!} \prod_{k=1}^{K} b(\mathbf{r}; \boldsymbol{\theta}), \end{aligned} \tag{3.9}$$

where the last step follows since $\int_{det} \mathrm{d}^2 r\, b(\mathbf{r}; \boldsymbol{\theta})$ is the total mean number of detected photons, $\bar{K}(\boldsymbol{\theta})$. The log-likelihood is

$$\ln \mathrm{pr}(\mathbf{G}|\boldsymbol{\theta}) = -\bar{K}(\boldsymbol{\theta}) - \ln K! + \sum_{k=1}^{K} \ln b(\mathbf{r}_k; \boldsymbol{\theta}). \tag{3.10}$$

C. Electronic Noise

Electronic noise comes from electrons, and in any practical system a very large number of electrons contribute more or less independently. It therefore follows from the central-limit theorem that electronic noise is accurately described by Gaussian statistics. Moreover, if we consider a discrete array of individual detector elements with no electronic coupling from one element to another, then the noise in different elements is statistically independent. Finally, if we assume that the elements are identical, the noise is modeled as i.i.d. (independent and identically distributed) zero-mean Gaussian. The optical illumination creates a signal that does not have zero mean, but if we assume that all noise sources are independent of the illumination, the effect of the illumination is to shift the noise PDF. Thus the only place that the parameter $\boldsymbol{\theta}$ can enter into the PDF on the data is in its mean. The PDF for purely electronic noise (without any photonic contribution) is given by

$$\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) = \sum_{m=1}^{M} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{[g_m - \bar{g}_m(\boldsymbol{\theta})]^2}{2\sigma^2} \right], \tag{3.11}$$

and its logarithm is

$$\ln \mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) = -\frac{1}{2}M\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{m=1}^{M}[g_m - \bar{g}_m(\boldsymbol{\theta})]^2. \tag{3.12}$$

Of the various assumptions that enter into Eqs. (3.11) and (3.12), the one that is the most suspect in practice is that the detector elements are identical. The pixels in commercial CCD detectors, for example, have considerable variation in dark current and responsivity. Postacquisition digital processing can correct these effects on average by subtracting a measured dark-current map and dividing the result by a measured gain map, but these corrections do not produce a uniform variance in each element; in fact, they may increase the variance nonuniformity since a pixel with low response will be divided by a small gain factor. A more accurate approach would be to measure the variances after the corrections and express the PDF on the corrected data as

$$\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) = \prod_{m=1}^{M}\frac{1}{\sqrt{2\pi\sigma_m^2}}\exp\left[-\frac{[g_m - \bar{g}_m(\boldsymbol{\theta})]^2}{2\sigma_m^2}\right]. \tag{3.13}$$

The FIM corresponding to Eq. (3.13) is readily shown to be

$$F_{jk} = \sum_{m=1}^{M}\frac{1}{\sigma_m^2}\frac{\partial\bar{g}_m(\boldsymbol{\theta})}{\partial\theta_j}\frac{\partial\bar{g}_m(\boldsymbol{\theta})}{\partial\theta_k}. \tag{3.14}$$

As with Poisson data, the only dependence of the likelihood or the Fisher information on the parameter is through $\bar{g}_m(\boldsymbol{\theta})$.

D. Combined Poisson and Gaussian Noise

So far we have discussed Poisson and Gaussian noise as if only one or the other were present, but in practice both will contribute in most cases.

Suppose the mth detector element receives k_m photoelectrons in some 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

$$\mathrm{pr}(g_m|\boldsymbol{\theta}) = \sum_{k_m=1}^{\infty}\mathrm{pr}(g_m|k_m)Pr(k_m|\boldsymbol{\theta}), \tag{3.15}$$

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. If we assume that all detectors have the same noise variance and responsivity, we obtain[41]

$$\begin{aligned}\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) &= \prod_{m=1}^{M}\mathrm{pr}(g_m|\boldsymbol{\theta}) \\ &= \prod_{m=1}^{M}\frac{1}{\sqrt{2\pi\sigma^2}}\sum_{k_m=0}^{\infty}\exp\left[-\frac{(g_m - Rk_m)^2}{2\sigma^2}\right] \\ &\quad\times\exp[-\bar{k}_m(\boldsymbol{\theta})]\frac{[\bar{k}_m(\boldsymbol{\theta})]^{k_m}}{k_m!}.\end{aligned} \tag{3.16}$$

Note that the only dependence on $\boldsymbol{\theta}$ in this expression is through the means $\bar{k}_m(\boldsymbol{\theta})$, so $\mathrm{pr}(g_m|\boldsymbol{\theta})$ can also be written as $\mathrm{pr}[g_m|\bar{k}_m(\boldsymbol{\theta})]$.

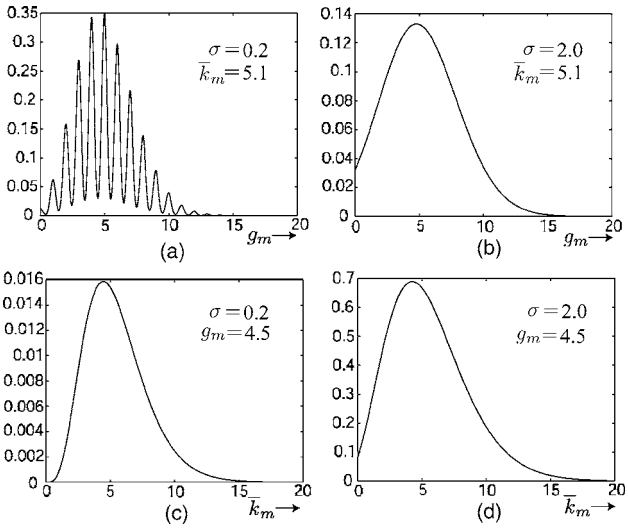

Fig. 2. Plots of $\mathrm{pr}[g_m|\bar{k}_m]$ for mixed Poisson and Gaussian noise: (a) and (b) show $\mathrm{pr}[g_m|\bar{k}_m]$ versus g_m for fixed $\bar{k}_m$; (c) and (d) show $\mathrm{pr}[g_m|\bar{k}_m]$ versus $\bar{k}_m$ for fixed g_m. Plots (a) and (c) are for small electronic noise (σ=0.2 in electron units), and plots (b) and (d) are for larger electronic noise (σ=2.0).

The dependence of $\mathrm{pr}[g_m|\bar{k}_m(\boldsymbol{\theta})]$ on g_m is illustrated in Figs. 2(a) and 2(b). The distinct peaks in Fig. 2(a) correspond to different integer numbers of detected photons. Figures 2(a) and 2(b) should not be confused with likelihoods; when $\mathrm{pr}[\mathbf{g}|k_m(\boldsymbol{\theta})]$ is plotted against $\bar{k}_m(\boldsymbol{\theta})$ for fixed g_m as in Figs. 2(c) and 2(d), a smooth unimodal likelihood results even when the variance of the electronic noise is small.

An exact expression for the FIM for combined Poisson and Gaussian noise is derived in Appendix A; a useful approximation is

$$F_{jk} \approx \sum_{m=1}^{M}\frac{R^2}{\sigma^2 + R^2\bar{k}_m(\boldsymbol{\theta})}\frac{\partial\bar{k}_m(\boldsymbol{\theta})}{\partial\theta_j}\frac{\partial\bar{k}_m(\boldsymbol{\theta})}{\partial\theta_k}, \tag{3.17}$$

where $\bar{k}_m(\boldsymbol{\theta})$ is the mean number of photoelectrons. This expression is exact for pure Gaussian noise or pure Poisson noise, and it is a good approximation for all values of $\bar{k}_m(\boldsymbol{\theta})$ so long as σ/R (the standard deviation of the electronic noise in photon units) is at least 0.5.

With combined Gaussian and Poisson noise, all you need to know to compute the FIM is $\bar{k}_m(\boldsymbol{\theta})$ (plus the detector characteristics R and σ^2, of course).

E. Detectors with Gain

Many detectors, including photomultipliers (PMTs), intensified CCDs, electron-multiplication CCDs, and avalanche photodiodes (APDs), have an internal gain mechanism to increase the level of the signal before subjecting it to electronic noise. Electron-multiplication CCDs are already being used in wavefront sensing, and arrays of APDs and multianode PMTs (essentially many PMTs in a common glass envelope) are also very promising for this application.

Two new features can arise in the stochastic data model for detectors with gain. The obvious one is that the gain process itself is noisy. A less-obvious effect is that in some cases the gain process can introduce correlations in the data values. In intensified CCDs or multianode PMTs, for example, the secondary electrons produced by a single primary photoelectron can spread over several neighboring output pixels.

Gain noise is no issue if the flux is low enough to allow thresholding and photon counting. The distribution of pulse heights is difficult to compute (see, for example, Saleh and Teich[42]), but it does not matter if the individual photons can be identified and counted.

Even spread of the secondaries to multiple pixels is not necessarily a problem at low photon flux; the electronics can be designed to recognize a cluster of pixels arising from a single primary event and to assign the event to a single pixel by some algorithm.[43] If these measures are taken (which they virtually never are), the output statistics remain rigorously uncorrelated Poisson[6] in spite of the gain noise and charge spread.

At the opposite extreme, if the primary photon flux is high and the detector simply integrates all of the charge at each pixel, then the effect of the gain noise in the absence of charge spread is mainly to increase the variance by a factor studied by Burgess[44] and Swank.[45] The case of amplification with spread has been studied by Rabbani and others.[46,47] For a review of this work, see Barrett and Myers,[6] Chap. 12. The outcome of these studies is easy to summarize if the mean number of primary photons per pixel is high; in that case we can invoke the central-limit theorem to say that the resulting overall PDF is multivariate Gaussian. The covariance matrix can be determined theoretically from the work cited above, or it can be measured for a particular detector. An important simplification in practice is that the correlations arising from charge spread will have short range, if they occur at all, so the covariance matrix will be diagonally dominant.

The intermediate case where the mean number of primary interactions per pixel is not low enough to permit identification of the signals from individual photons, yet not high enough that the central-limit theorem is valid, is just beginning to receive scrutiny.[48]

F. PDF and Likelihood for Correlated Gaussian Noise

As we have seen, there are several possible situations in which the data provided by a WFS can be described as correlated Gaussian. In Subsection 3.E, we discussed correlations arising from charge spread in certain detectors with gain. Without charge spread, the data will be inherently uncorrelated, at least if we define the correlation with respect to the conditional PDF $\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ includes all parameters that can affect the mean data. When we use some subset of these parameters, however, it often turns out that there are correlations induced by the parameters we choose to leave out (see Subsection 5.A). Finally, as we shall see in Appendix D, computation of centroids or other derived parameters usually results in correlations. In all of these cases, it may turn out that a more realistic data PDF is the correlated multivariate normal Gaussian.

A general multivariate normal PDF has the form:

$$\mathrm{pr}(\mathbf{g}) = [(2\pi)^M \det(\mathbf{K})]^{-1/2} \exp\left[-\tfrac{1}{2}(\mathbf{g}-\bar{\mathbf{g}})^t \mathbf{K}^{-1}(\mathbf{g}-\bar{\mathbf{g}})\right], \tag{3.18}$$

where $\bar{\mathbf{g}}$ is the mean vector and $\mathbf{K}$ is the covariance matrix of $\mathbf{g}$. The most general likelihood function is obtained by letting the mean and covariance both be functions of $\boldsymbol{\theta}$:

$$\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta}) = [(2\pi)^M \det[\mathbf{K}(\boldsymbol{\theta})]]^{-1/2} \times \exp\left\{-\tfrac{1}{2}[\mathbf{g}-\bar{\mathbf{g}}(\boldsymbol{\theta})]^t [\mathbf{K}(\boldsymbol{\theta})]^{-1} [\mathbf{g}-\bar{\mathbf{g}}(\boldsymbol{\theta})]\right\}. \tag{3.19}$$

4. PARAMETERIZATION

As in Subsection 2.C, here we shall denote the parameters we want to estimate by the $N\times 1$ vector $\boldsymbol{\alpha}$, but we must recognize that this parameter set is seldom sufficient either to specify the wavefront fully or to completely describe the PDF of the data. In this section we look at some choices for $\boldsymbol{\alpha}$, what they imply for our representation of the wavefront, and how they have to be augmented to get the full parameter set $\boldsymbol{\theta}$ that describes the data.

A. Wavefront Representations

Suppose the wave incident on the WFS has the form $\exp[ikW(\mathbf{r})]$, where $\mathbf{r}=(x,y)$ and $k=2\pi/\lambda$. Let $\{\gamma_n,\ n=1,\ldots,\infty\}$ denote an infinite set of parameters that can be used to express an arbitrary wavefront exactly as

$$W(\mathbf{r}) = \sum_{n=1}^{\infty} \gamma_n u_n(\mathbf{r}), \tag{4.1}$$

where the set $\{u_n(\mathbf{r})\}$ is some orthonormal basis (e.g., Zernike polynomials). It is safe to say that we are never interested in estimating the full wavefront or the infinite-dimensional vector $\boldsymbol{\gamma}$.

Sometimes we are interested in the N lowest-order terms in Eq. (4.1) for their own sake. In ophthalmology, for example, we might want to estimate the first N Zernike coefficients in order to use them for the task of planning laser surgery. In that case a reasonable choice for the parameters of interest would be $\alpha_n=\gamma_n,\ n=1,\ldots,N$.

In AO, however, the usual objective is to determine the signals to be applied to the actuators of a deformable mirror. The possible phase functions that can be produced by a deformable mirror are assumed to be linear combinations of its influence functions $\{\psi_n(\mathbf{r}),\ n=1,\ldots N\}$, where N is the number of actuators. With this consideration in mind, we can write Eq. (4.1) in the form

$$W(\mathbf{r}) = \sum_{n=1}^{N} \alpha_n \psi_n(\mathbf{r}) + \Delta W(\mathbf{r}). \tag{4.2}$$

The $N\times 1$ vector $\boldsymbol{\alpha}$ is what is needed for mirror control and hence a reasonable choice of parameters to estimate, and $\Delta W(\mathbf{r})$ will be referred to as the *residual*. If the coefficients $\{\alpha_n\}$ are chosen by least-squares (LS) fitting, the residual is orthogonal to the sum and Eq. (4.2) is an orthogonal decomposition of the wavefront.

Another way of representing a wavefront is to divide it into regions (subapertures), approximate the wavefront

over each region by a small set of known functions that are zero outside the region, and then append a residual as in Eq. (4.2) to make the expansion exact. The coefficients in the regional representation can then be estimated, not for their own intrinsic interest, but so they can be used in a subsequent estimation of the mirror-mode coefficients $\boldsymbol{\alpha}$.

As an example, consider a representation in terms of local tilts. Suppose the jth region ($j=1,\ldots,J$) is centered at $\mathbf{r}=\mathbf{r}_j$, or equivalently $x=x_j$ and $y=y_j$. Let the region itself be defined by a support function $S_j(\mathbf{r})$, which is unity for $\mathbf{r}$ inside the region and zero outside. We assume that all regions are identical, so $S_j(\mathbf{r})=S(\mathbf{r}-\mathbf{r}_j)$, and we assume that different regions do not overlap. Local tilt functions in the x and y directions can now be defined by

$$\chi_k(\mathbf{r}) = \begin{cases} S(\mathbf{r}-\mathbf{r}_j)(x-x_j) & j=(k+1)/2 \quad \text{if } k \text{ odd} \\ S(\mathbf{r}-\mathbf{r}_j)(y-y_j) & j=k/2 \quad \text{if } k \text{ even} \end{cases}. \tag{4.3}$$

These functions are orthogonal for square apertures, but they are not normalized.

With the tilt functions, a representation similar to Eq. (4.2) can be given as

$$W(\mathbf{r}) = \sum_{k=1}^{2J} \tau_k \chi_k(\mathbf{r}) + \delta W(\mathbf{r}). \tag{4.4}$$

This representation is particularly useful if the region is small enough (e.g., much smaller than the Fried parameter in the atmospheric case) since then it may be a good approximation to say that the wavefront in the region is described completely by its tilts and pistons. The tilts are accounted for by the sum in Eq. (4.4), and the pistons are contained in $\delta W(\mathbf{r})$. For a square aperture, the local piston is orthogonal to the tilt function so Eq. (4.4), like Eq. (4.2), is an orthogonal decomposition of the wavefront.

B. Nuisance Parameters

There are two distinct classes of nuisance parameters in wavefront sensing: intrinsic nuisance parameters related to the wavefront expansion itself and extrinsic nuisance parameters that arise from other sources.

Examples of extrinsic nuisance parameters include the brightness of the guide star, length of the sodium column when a laser guide star is used, level and distribution of background light, and scintillation effects. Which of these we need to consider depends on the application and the data-acquisition system; in Section 5 we shall consider brightness of the guide star and background light level as examples.

Intrinsic nuisance parameters are the ones needed to represent the residual in Eq. (4.2) or (4.4). Since the residual is an infinite-dimensional function (technically a vector in the Hilbert space $\mathbb{L}_2(\mathbb{R}^2)$), it might appear that an infinite set of parameters would be needed, but not all components of the residual influence the data.

One way to parameterize the residual is to recognize that the sum in Eq. (4.2) or (4.4) defines a vector in a subspace of $\mathbb{L}_2(\mathbb{R}^2)$. Following terminology introduced by Paxman,[49] we can refer to this subspace as *interest space* and to its orthogonal complement as *indifference space*. If we are interested in estimating the signals needed to control a deformable mirror as in Eq. (4.2), for example, the mirror influence functions form a (nonorthogonal) basis for interest space, and all functions in indifference space are orthogonal to all influence functions.

We can define an orthonormal basis $\{\Xi_k(\mathbf{r})\}$ for indifference space by use of projection operators (see Barrett and Myers[6] for details), and then we can represent the residual as

$$\Delta W(\mathbf{r}) = \sum_{k=1}^{\infty} \beta_k^{int} \Xi_k(\mathbf{r}). \tag{4.5}$$

Though this sum is infinite, only a finite subset of the terms, say K of them, will influence the data significantly, and we can use those coefficients to define a $K\times 1$ vector $\boldsymbol{\beta}^{int}$ that describes the intrinsic nuisance parameters.

Table 1. Vectors Relevant to Wavefront Sensing

Vector	Meaning	Dimension
$\mathbf{g}$	Raw data (photodetector outputs)	$M\times 1$
$\boldsymbol{\alpha}$	Parameters of interest (e.g., mirror modes)	$N\times 1$
$\boldsymbol{\beta}^{int}$	Intrinsic nuisance parameters	$K\times 1$
$\boldsymbol{\beta}^{ext}$	Extrinsic nuisance parameters	$L\times 1$
$\boldsymbol{\beta}$	All nuisance parameters	$(K+L)\times 1$
$\boldsymbol{\theta}$	All parameters that influence data	$P\times 1$, $(P=N+K+L)$
$\boldsymbol{\gamma}$	Parameters in exact wavefront representation	$\infty\times 1$
$\boldsymbol{\tau}$	Coefficients of local tilt functions in J subapertures	$2J\times 1$

C. Summary of Parameters

The vectors that will be needed in Section 5 are summarized in Table 1.

5. PRACTICAL LIKELIHOOD FUNCTIONS AND FISHER INFORMATION MATRICES

The goal of this section is to show how the general principles discussed above can be used to construct practical likelihood functions and FIMs. Emphasis in this section will be on the problem of directly estimating the mirror modes without the intermediary of the reconstruction stage in Fig. 1, but in Section 6 we consider the more common problem of estimating local tilts from Shack–Hartmann data.

Any of the likelihood functions developed in this section can be used for MAP estimation as well, provided one has a meaningful prior on the parameters to be estimated.

A. General Considerations on Nuisance Parameters

The first decision we have to make in constructing a practical likelihood function is what to do about intrinsic and extrinsic nuisance parameters. The possibilities were enumerated in Subsection 2.C; which option we use depends in large part on the dimensionality of the nuisance parameter.

To be explicit, consider two specific extrinsic nuisance parameters in astronomical wavefront sensing: the brightness of the guide star and the average sky background. These two numbers form the components of a 2×1 extrinsic nuisance parameter vector. Both can affect the mean data strongly, so they should not be ignored [option (1) in Subsection 2.C]. Both vary significantly with site, guide star chosen, and position in the sky, so typical values [option (2)] would not be reliable, and prior PDFs [option (5)] would be broad and relatively uninformative. As we shall see below, however, both parameters can be estimated from the same data as used to estimate the wavefront parameters [option (3)] or from some expanded data set [option (4)], and these would have to be the recommended options.

Often, however, intrinsic and extrinsic nuisance parameters require high-dimensional parameter vectors. The sky background, for example, might be a complicated spatial distribution rather than just a single number, and many different modes can contribute to the intrinsic nuisance parameter $\boldsymbol{\beta}^{int}$. In these cases any attempt to estimate all components will increase the dimension and condition number of the FIM and thereby increase the CRB on the parameters of interest. (For a proof of this statement, see Barrett and Myers,[6] Sec. 13.3.8.). If the number of nuisance parameters is larger than the number of measurements, the FIM is singular and the CRB is infinite.

With high-dimensional nuisance parameters, therefore, the only remaining options are to ignore them [option (1)] or to marginalize over them [option (5)]. To reiterate a point from Subsection 2.C, marginalization is optimal in terms of risk if a meaningful prior is known.

B. Marginalizing Intrinsic Nuisances

If we are interested in estimating $\boldsymbol{\alpha}$ from a data set $\mathbf{g}$ by ML (or MAP) methods, we need the likelihood $\mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha})$. What we know from Section 3, however, is $\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})$ or $\mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha},\boldsymbol{\beta})$. If we want to marginalize over all nuisance parameters, we need

$$\mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha}) = \int \mathrm{d}^{K+L}\beta\, \mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha},\boldsymbol{\beta})\mathrm{pr}(\boldsymbol{\beta}|\boldsymbol{\alpha}), \tag{5.1}$$

and if we want to marginalize over just the intrinsic nuisance parameters and estimate the extrinsic ones, we need

$$\mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha},\boldsymbol{\beta}^{ext}) = \int \mathrm{d}^{K}\beta^{int}\, \mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha},\boldsymbol{\beta}^{ext},\boldsymbol{\beta}^{int})\mathrm{pr}(\boldsymbol{\beta}^{int}|\boldsymbol{\alpha}). \tag{5.2}$$

Note that we do not write $\mathrm{pr}(\boldsymbol{\beta}^{int}|\boldsymbol{\alpha},\boldsymbol{\beta}^{ext})$ in Eq. (5.2) because there is no apparent way that extrinsic parameters like guide-star brightness and sky background can influence the wavefront being sensed.

In both Eqs. (5.1) and (5.2), a conditional prior on $\boldsymbol{\beta}$ is needed, and in keeping with the spirit of this paper, it has to be a prior with experimental justification.

In astronomy, there is a large body of experimental evidence supporting the Kolmogorov theory of atmospheric turbulence. Central to that theory is the assumption that phase perturbations are zero-mean Gaussian random processes, so the coefficient of any term in any linear representation of a wavefront must be a Gaussian random variable. We may therefore safely take $\mathrm{pr}(\boldsymbol{\beta}^{int})$ as a K-dimensional zero-mean multivariate normal density. What we need in Eq. (5.2), however, is $\mathrm{pr}(\boldsymbol{\beta}^{int}|\boldsymbol{\alpha})$ rather than $\mathrm{pr}(\boldsymbol{\beta}^{int})$, and the dependence on $\boldsymbol{\alpha}$ is a complication since that is the main parameter we want to estimate.

There are two ways we can justify replacing $\mathrm{pr}(\boldsymbol{\beta}^{int}|\boldsymbol{\alpha})$ in Eq. (5.2) with a multivariate normal independent of $\boldsymbol{\alpha}$. The obvious one is simply to assume that $\boldsymbol{\beta}^{int}$ is independent of $\boldsymbol{\alpha}$. A more subtle approach is to recognize that in a closed-loop system where $\boldsymbol{\alpha}$ represents the coefficients of the mirror modes, the effect of the AO system is to drive $\boldsymbol{\alpha}$ close to zero. We can formalize this notion by the closed-loop approximation:

$$\mathrm{pr}(\boldsymbol{\beta}^{int}|\boldsymbol{\alpha}) \approx \mathrm{pr}(\boldsymbol{\beta}^{int}|\boldsymbol{\alpha}=\mathbf{0}). \tag{5.3}$$

It is shown in Appendix B that $\mathrm{pr}(\boldsymbol{\beta}^{int}|\boldsymbol{\alpha}=0)$ is itself a zero-mean multivariate normal of the form

$$\mathrm{pr}(\boldsymbol{\beta}^{int}|\boldsymbol{\alpha}=\mathbf{0}) = \mathcal{N}\exp\Big[-\tfrac{1}{2}(\boldsymbol{\beta}^{int})^t\mathbf{C}^{-1}(\boldsymbol{\beta}^{int})\Big], \tag{5.4}$$

where $\mathcal{N}=[(2\pi)^K\det(\mathbf{C})]^{-1/2}$ and $\mathbf{C}$ is a covariance matrix known as a Schur complement; if $\boldsymbol{\beta}^{int}$ and $\boldsymbol{\alpha}$ were uncorrelated, $\mathbf{C}$ would be just the covariance matrix of $\boldsymbol{\beta}^{int}$. With Eqs. (5.3) and (5.4), the desired likelihood function [Eq. (5.2)] becomes

$$\mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha},\boldsymbol{\beta}^{ext}) \approx \mathcal{N}\int \mathrm{d}^{K}\beta^{int}\, \mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha},\boldsymbol{\beta}^{ext},\boldsymbol{\beta}^{int}) \times\exp\Big[-\tfrac{1}{2}(\boldsymbol{\beta}^{int})^t\mathbf{C}^{-1}(\boldsymbol{\beta}^{int})\Big]. \tag{5.5}$$

To proceed, we must choose a form for the likelihood conditional on all relevant parameters, $\mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha},\boldsymbol{\beta}^{ext},\boldsymbol{\beta}^{int})$. The simplest choice is the i.i.d. normal model presented in Subsection 3.C. Using Eqs. (3.11) and (5.5), we can write

$$\mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha},\boldsymbol{\beta}^{ext}) \approx \mathcal{N}'\int_{\infty} \mathrm{d}^{K}\beta^{int} \times\exp\left[-\sum_{m=1}^{M}\frac{[g_m-\bar{g}_m(\boldsymbol{\alpha},\boldsymbol{\beta}^{ext},\boldsymbol{\beta}^{int})]^2}{2\sigma^2}\right] \times\exp\Big[-\tfrac{1}{2}(\boldsymbol{\beta}^{int})^t\mathbf{C}^{-1}(\boldsymbol{\beta}^{int})\Big], \tag{5.6}$$

where the integral runs from $-\infty$ to ∞ over all K variables and $\mathcal{N}'=\mathcal{N}(2\pi\sigma^2)^{-M/2}$. This integral would be the convolution of two Gaussians, immediately yielding another Gaussian, except that $\boldsymbol{\beta}^{int}$ enters into the first factor in the integrand in a complicated way through the mean $\bar{g}_m(\boldsymbol{\alpha},\boldsymbol{\beta}^{ext},\boldsymbol{\beta}^{int})$; we can fix this problem by assuming that the effect of $\boldsymbol{\beta}^{int}$ is small, performing a Taylor expansion of the mean, and retaining only the first two terms. Details are given in Appendix B, where it is shown that

$$\mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha},\boldsymbol{\beta}^{ext}) \approx \mathcal{N}''\exp\Big\{-\tfrac{1}{2}[\mathbf{g}-\bar{\mathbf{g}}(\boldsymbol{\alpha},\boldsymbol{\beta}^{ext},\mathbf{0})]^t \times\mathbf{K}_{tot}^{-1}[\mathbf{g}-\bar{\mathbf{g}}(\boldsymbol{\alpha},\boldsymbol{\beta}^{ext},\mathbf{0})]^t\Big\}, \tag{5.7}$$

where $\mathcal{N}''=[(2\pi)^M\det(\mathbf{K}_{tot})]^{-1/2}$ and

$$\mathbf{K}_{tot} \equiv \sigma^2 \mathbf{I} + \mathbf{ACA}^t, \tag{5.8}$$

with **A** being a matrix defined in the appendix. Note that the fact that $\bar{g}_m(\boldsymbol{\alpha}, \boldsymbol{\beta}^{ext}, \boldsymbol{\beta}^{int})$ is evaluated at $\boldsymbol{\beta}^{int} = \mathbf{0}$ does not mean that the unwanted modes are being set to zero; rather it comes from the assumption that $\boldsymbol{\beta}^{int}$ has zero mean and that excursions about the mean are small enough to allow a first-order Taylor expansion.

To the first order, Eq. (5.7) shows that the likelihood after marginalizing over the intrinsic nuisance parameters is a multivariate normal with mean determined without any consideration of the nuisance parameters. To this order, the only effect of the unwanted modes is to add a new, nondiagonal term to the covariance matrix. This result generalizes easily to include readout noise that varies from detector to detector, gain noise, and even photon noise so long as the Poisson can be approximated by a Gaussian.

In practice, neither **C** nor **A** is known, but it is straightforward to simulate realizations of Kolmogorov turbulence, either fully digitally or with a spatial light modulator, and to find a sample covariance matrix that is an experimental approximation to $\mathbf{ACA}^t$. The matrix inversion required in Eq. (5.7) can then be performed by methods described in Chap. 14 of Barrett and Myers,[6] even if the sample covariance matrix is not full rank.

To summarize this subsection, we have seen that there are several possible approaches to choosing a prior with which to marginalize over the nuisance parameters. In the view of the authors, the final justification for making this choice will have to come from a meaningful, task-based performance assessment of the overall AO system.[35]

C. Poisson Data with Negligible Intrinsic Nuisances

Sometimes we can get away with the assumption that there are no intrinsic nuisance parameters. In Shack–Hartmann sensors with relatively small subapertures, for example, it is probably valid to neglect aberrations other than piston and tilt; piston does not affect the data, and tilt is what we want to estimate, so there are no intrinsic nuisance parameters.

If there are no significant intrinsic nuisance parameters and we choose to estimate the extrinsic ones, then all of the likelihood functions and FIMs derived in Section 3 are immediately applicable, just by identifying

$$\boldsymbol{\theta} = \begin{pmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta}^{ext} \end{pmatrix}.$$

In particular, for pure Poisson data, the log-likelihood is given by Eq. (3.2), which we can rewrite as

$$\ln \Pr(\mathbf{g}|\boldsymbol{\alpha}, \boldsymbol{\beta}^{ext}) = \sum_{m=1}^{M} \{-\bar{g}_m(\boldsymbol{\alpha}, \boldsymbol{\beta}^{ext}) + g_m \ln[\bar{g}_m(\boldsymbol{\alpha}, \boldsymbol{\beta}^{ext})]\}. \tag{5.9}$$

The term $\ln g_m!$ has been dropped since it is independent of the parameters and hence does not affect the likelihood [$\Pr(\mathbf{g}|\boldsymbol{\theta})$ regarded as a function of $\boldsymbol{\theta}$ for fixed $\mathbf{g}$].

Consider the case where the extrinsic nuisance parameter is only the brightness of the guide star (or other point source), denoted I_0. In that case we can express the mean data as

$$\bar{g}_m(\boldsymbol{\alpha}, I_0) = I_0 f_m(\boldsymbol{\alpha}), \tag{5.10}$$

where $f_m(\boldsymbol{\alpha})$ is a characteristic of the individual detector element, defined in such a way that $I_0 f_m(\boldsymbol{\alpha})$ is the mean number of photons detected by the mth element when the wavefront is fully described by the vector $\boldsymbol{\alpha}$. The log-likelihood is now given by

$$\ln \Pr(\mathbf{g}|\boldsymbol{\alpha}, I_0) = -I_0 \sum_{m=1}^{M} f_m(\boldsymbol{\alpha}) + \sum_{m=1}^{M} g_m \ln[f_m(\boldsymbol{\alpha})] + N_{tot} \ln(I_0), \tag{5.11}$$

where $N_{tot} \equiv \Sigma_{m=1}^{M} g_m$ is the total number of detected photons.

1. Fisher Information with One Nuisance Parameter

If $\boldsymbol{\alpha}$ is an $N \times 1$ vector and the only nuisance parameter is the guide-star brightness, then the FIM is $(N+1) \times (N+1)$. The derivatives needed in the FIM are

$$\frac{\partial}{\partial \alpha_n} \ln \Pr(\mathbf{g}|\boldsymbol{\alpha}, I_0) = \sum_{m=1}^{M} \left[\frac{g_m - \bar{g}_m(\boldsymbol{\alpha}, I_0)}{f_m(\boldsymbol{\alpha})} \right] \frac{\partial f_m(\boldsymbol{\alpha})}{\partial \alpha_n}, \tag{5.12}$$

$$\frac{\partial}{\partial I_0} \ln \Pr(\mathbf{g}|\boldsymbol{\alpha}, I_0) = \frac{1}{I_0} \sum_{m=1}^{M} [g_m - \bar{g}_m(\boldsymbol{\alpha}, I_0)]. \tag{5.13}$$

The statistical average needed in the FIM is

$$\left\langle [g_m - \bar{g}_m(\boldsymbol{\alpha}, I_0)][g_{m'} - \bar{g}_{m'} - (\boldsymbol{\alpha}, I_0)] \right\rangle_{\mathbf{g}|\boldsymbol{\alpha}, I_0} = \bar{g}_m(\boldsymbol{\alpha}, I_0) \delta_{mm'}, \tag{5.14}$$

and the elements of the FIM are found to be

$$F_{nn'} = I_0 \sum_{m=1}^{M} \frac{1}{f_m(\boldsymbol{\alpha})} \frac{\partial f_m(\boldsymbol{\alpha})}{\partial \alpha_n} \frac{\partial f_m(\boldsymbol{\alpha})}{\partial \alpha_{n'}} \quad (n, n' \le N), \tag{5.15}$$

$$F_{n,N+1} = F_{N+1,n} = \sum_{m=1}^{M} \frac{\partial f_m(\boldsymbol{\alpha})}{\partial \alpha_n} \quad (n \le N), \tag{5.16}$$

$$F_{N+1,N+1} = \frac{\bar{N}_{tot}}{I_0^2} = \frac{1}{I_0} \sum_{m=1}^{M} f_m(\boldsymbol{\alpha}), \tag{5.17}$$

We see, therefore, that the FIM for this problem is a partitioned matrix with the structure

$$\mathbf{F} = \begin{bmatrix} \mathbf{A}_{N \times N} & \mathbf{B}_{N \times 1} \\ \mathbf{B}^t_{1 \times N} & \mathbf{C}_{1 \times 1} \end{bmatrix}, \tag{5.18}$$

where the elements of **A** [given by Eq. (5.15)] scale as I_0, the elements of **B** [given by Eq. (5.16)] are independent of I_0, and **C** is proportional to $1/I_0$.

2. Inclusion of Sky Background

Now we consider an additional nuisance, the sky background treated as a uniform incoherent source. This additional radiation does not spoil the Poisson assumptions, but instead modifies the mean data with an additional term. If each detector receives the same amount of sky radiation on average, then Eq. (5.10) becomes

$$\bar{g}_m(\boldsymbol{\alpha}, I_0, b) = I_0 f_m(\boldsymbol{\alpha}) + b, \tag{5.19}$$

where the scalar b, defined as the mean number of detected background photons per pixel, is now one additional nuisance parameter. If dark current is significant, its effect can also be included in b.

With two nuisance parameters, the log-likelihood Eq. (5.11) becomes

$$\ln \Pr(\mathbf{g}|\boldsymbol{\alpha}, I_0, b) = -I_0 \sum_{m=1}^{M} f_m(\boldsymbol{\alpha}) - Mb + \sum_{m=1}^{M} g_m \ln[I_0 f_m(\boldsymbol{\alpha}) + b]. \tag{5.20}$$

The FIM is now $(N+2)\times(N+2)$, and the derivatives needed for its computation are

$$\frac{\partial}{\partial \alpha_n} \ln \Pr(\mathbf{g}|\boldsymbol{\alpha}, I_0, b) = I_0 \sum_{m=1}^{M} \left[\frac{g_m - \bar{g}_m(\boldsymbol{\alpha}, I_0, b)}{\bar{g}_m(\boldsymbol{\alpha}, I_0, b)} \right] \frac{\partial f_m(\boldsymbol{\alpha})}{\partial \alpha_n}, \tag{5.21}$$

$$\frac{\partial}{\partial I_0} \ln \Pr(\mathbf{g}|\boldsymbol{\alpha}, I_0, b) = \sum_{m=1}^{M} [g_m - \bar{g}_m(\boldsymbol{\alpha}, I_0, b)] \frac{f_m(\boldsymbol{\alpha})}{I_0 f_m(\boldsymbol{\alpha}) + b}, \tag{5.22}$$

$$\frac{\partial}{\partial b} \ln \Pr(\mathbf{g}|\boldsymbol{\alpha}, I_0, b) = \sum_{m=1}^{M} \left[\frac{g_m - \bar{g}_m(\boldsymbol{\alpha}, I_0, b)}{\bar{g}_m(\boldsymbol{\alpha}, I_0, b)} \right]. \tag{5.23}$$

The elements of $\mathbf{F}$ can now be computed with the help of a slight generalization of Eq. (5.14).

D. Maximum-Likelihood Estimation from Gaussian Measurements

Subsection 5.C dealt with purely Poisson noise, but we saw earlier that there are several situations in which the Poisson model is incorrect. Electronic readout noise and gain noise are continuous random variables and hence not Poisson, and we saw in Subsection 5.B that marginalizing over unwanted wavefront modes can yield a multivariate Gaussian likelihood.

It is well known that ML estimation with Gaussian data is basically LS fitting. If the mean data are linear functions of the parameters to be estimated, then ML estimation is the same as linear regression, with the regression function being the negative of the log-likelihood. The ML solution in this case is obtained by matrix inversion or pseudoinversion.[6] In wavefront sensing and many other applications, however, the mean data depend nonlinearly on the parameters, so no linear method will deliver ML estimates.

1. Independent Gaussian Measurements

A general likelihood for statistically independent Gaussian measurements is given in Eq. (3.13). If we allow the variance to depend on $\boldsymbol{\theta}$ for generality, the corresponding log-likelihood boils down to

$$\ln \operatorname{pr}(\mathbf{g}|\boldsymbol{\theta}) = -\tfrac{1}{2} \sum_{m=1}^{M} \frac{[g_m - \bar{g}_m(\boldsymbol{\theta})]^2}{\sigma_m^2(\boldsymbol{\theta})} + \text{constant}. \tag{5.24}$$

Because of the leading minus sign, maximizing the log-likelihood is the same thing as minimizing a weighted norm of the difference between the measured data vector $\mathbf{g}$ and the predicted mean data $\bar{\mathbf{g}}(\boldsymbol{\theta})$. ML estimation from independent Gaussian data is a nonlinear regression.

2. Correlated Gaussian Measurements

Detectors with gain may deliver inherently correlated Gaussian data, and marginalizing over nuisance parameters may induce correlations even when the detectors themselves do not. The log-likelihood in these cases is given by

$$\ln \operatorname{pr}(\mathbf{g}|\boldsymbol{\theta}) = -\tfrac{1}{2} \sum_{m=1}^{M} \sum_{m'=1}^{M} [g_m - \bar{g}_m(\boldsymbol{\theta})][\mathbf{K}^{-1}]_{mm'}[g_{m'} - \bar{g}_{m'}(\boldsymbol{\theta})] = -\tfrac{1}{2}[\mathbf{g} - \bar{\mathbf{g}}(\boldsymbol{\theta})]^t \mathbf{K}^{-1} [\mathbf{g} - \bar{\mathbf{g}}(\boldsymbol{\theta})], \tag{5.25}$$

where $\mathbf{K}$ is a covariance matrix which, in the most general case, can depend on $\boldsymbol{\theta}$.

6. APPLICATION TO A SHACK–HARTMANN SENSOR

Though the likelihood models developed above are applicable to any wavefront sensor, the familiar Shack–Hartmann sensor provides an instructive example. In its simplest form, a Shack–Hartmann sensor consists of an array of lenslets in, say, the plane $z=0$, and an array of photodetectors in a parallel plane, $z=z_0$ (where z_0 is not necessarily the focal length of the lenslets). The data from the entire detector array can, in principle, be used to estimate the full set of parameters of interest $\boldsymbol{\alpha}$, but in practice a subset of the data associated with a single lenslet is used to estimate local tilts, which are then used to estimate $\boldsymbol{\alpha}$ in a separate reconstruction step. In this section we first look at the conventional problem of estimation of local tilts and then discuss the application of likelihood principles to estimation of $\boldsymbol{\alpha}$.

A. Estimation of Local Tilts from Poisson Data

If the geometry in a Shack–Hartmann sensor is chosen so that radiation passing through one lenslet falls only on one subset of the detector pixels, then the local wavefront parameters for each lenslet can be estimated independently of those for other lenslets. Moreover, if the wave over one lenslet is well described as a pure tilt, then there are no intrinsic nuisance parameters, and the likelihood functions given in Subsections 5.C and 5.D are applicable if we simply replace the general parameter $\boldsymbol{\alpha}$ with the 2D tilt vector $\boldsymbol{\tau}$ for the lenslet of interest.

In particular, if the noise is Poisson and the unknowns are the guide-star brightness and two components of the

local tilt, then the log-likelihood, given by Eq. (5.11), is specified by the set of functions $\{f_m(\boldsymbol{\tau})\}$, where the index m now runs over only those detector elements that receive radiation from the particular lenslet. For a normally incident plane wave in a Shack–Hartmann sensor, the lenslet produces an irradiance distribution on the detector plane (a "spot") denoted by $s(\mathbf{r})$. If z_0 is the focal length of the lenslet, then $s(\mathbf{r})$ is the squared modulus of the (suitably scaled) Fourier transform of the pupil function, but in general it can also be a defocused image of the pupil. In either case, the effect of a pure tilt is to shift the spot, and the mean output of the mth detector element is obtained by multiplying the irradiance by the responsivity function of that element, $d_m(\mathbf{r})$, and integrating

$$f_m(\boldsymbol{\tau}) = \int_{\infty} \mathrm{d}^2 r\, d_m(\mathbf{r}) s(\mathbf{r} - z_0 \boldsymbol{\tau}). \tag{6.1}$$

The units are again chosen so that $I_0 f_m(\boldsymbol{\tau})$ is the mean number of photons from the guide star detected in element m. Thus $f_m(\boldsymbol{\tau})$ is the mean response of the detector element as a function of the shift of the spot.

1. Some Simplifying Assumptions

A common assumption made in analyzing Shack–Hartmann sensors is that there is no light loss as the spot shifts, so that

$$\sum_{m=1}^{M_1} f_m(\boldsymbol{\tau}) \equiv f_{tot} = \text{constant}, \tag{6.2}$$

where M_1 is the number of detector elements associated with a particular lenslet and $\boldsymbol{\tau}$ is the 2D vector of x and y tilts over that lenslet. The assumption in Eq. (6.2) is valid if (a) there are no gaps between detector elements; (b) the responsivity of all detector elements is the same; (c) obliquity and other angular factors are neglected; (d) the spot does not fall off the area of the detector associated with the lenslet; and (e) that detector area does not receive light from adjacent lenslets. With these restrictive physical assumptions *and* the assumptions of pure Poisson noise, no intrinsic nuisance parameters and no sky background, the log-likelihood from Eq. (5.11) becomes

$$\ln \Pr(\mathbf{g}|\boldsymbol{\tau}, I_0) = -f_{tot} I_0 + \sum_{m=1}^{M_1} g_m \ln[f_m(\boldsymbol{\tau})] + N_{tot} \ln(I_0), \tag{6.3}$$

where $\boldsymbol{\tau}$ is now a 2D vector specifying the x and y components of tilt over that lenslet.

Equation (6.3) is the form of the log-likelihood used most commonly in the literature on wavefront sensing, though it is also common to go further and consider a very large number of small detector elements so that $d_m(\mathbf{r})$ can be treated as a delta function.

One advantage of assumption (6.2) is that the FIM becomes block diagonal since

$$\sum_{m=1}^{M_1} \frac{\partial f_m(\boldsymbol{\tau})}{\partial \tau_n} = \frac{\partial}{\partial \tau_n} \sum_{m=1}^{M_1} f_m(\boldsymbol{\tau}) = 0. \tag{6.4}$$

Thus, as shown by Eq. (5.16), the off-diagonal blocks $\mathbf{B}$ in Eq. (5.18) vanish, and the CRBs on $\boldsymbol{\tau}$ and I_0 are readily computed.

A consequence of the block-diagonal FIM is that the CRB on the estimates of the parameters of interest, $\boldsymbol{\tau}$, is obtained just by inverting the $\mathbf{A}$ block in Eq. (5.18). Therefore it is the same as if $\mathbf{C}$ were not present, and there is no penalty in the performance bound for including I_0 in the parameter list.

Another consequence of model (6.3) and the block-diagonal FIM is that I_0 and $\boldsymbol{\tau}$ can be estimated separately. The ML estimate of I_0 is obtained by setting $\partial[\ln \Pr(\mathbf{g}|\boldsymbol{\tau}, I_0)]/\partial I_0$ as given by Eq. (5.13) to zero and by using Eq. (5.10); the result is

$$\hat{I}_0 = \frac{\sum_{m=1}^{M_1} g_m}{\sum_{m=1}^{M_1} f_m(\boldsymbol{\tau})} = \frac{N_{tot}}{\sum_{m=1}^{M_1} f_m(\boldsymbol{\tau})}. \tag{6.5}$$

If Eq. (6.2) holds, the denominator is independent of the $\boldsymbol{\tau}$ and the guide-star brightness can be estimated independently of the tilts. The ML tilt estimates $\hat{\boldsymbol{\tau}}$ are then found by setting $\partial[\ln \Pr(\mathbf{g}|\boldsymbol{\tau}, I_0)]/\partial_{\tau_n}$ as given by Eq. (5.12) to zero. The result is

$$\sum_{m=1}^{M_1} \frac{g_m}{f_m(\boldsymbol{\tau})} \frac{\partial f_m(\boldsymbol{\tau})}{\partial \tau_n} = 0 \quad \text{when } \boldsymbol{\tau} = \hat{\boldsymbol{\tau}}. \tag{6.6}$$

This result does not require knowledge of the guide-star brightness, so we may as well ignore it; we emphasize, however, that this result requires that there be no light loss, no overlapping with adjacent lenslets, no sky background, and pure Poisson noise.

2. Joint Estimation of Tilts and Nuisance Parameters

If Eq. (6.2) does not hold or if there is a sky background, all parameters associated with a single subaperture must be estimated jointly. The derivative formulas are not particularly useful, and the best we can say is that the log-likelihood from Eq. (5.20) must be maximized:

$$-I_0 \sum_{m=1}^{M_1} f_m(\boldsymbol{\tau}) - Mb + \sum_{m=1}^{M_1} g_m \ln[I_0 f_m(\boldsymbol{\tau}) + b] = \text{maximum} \quad \text{at } \boldsymbol{\tau} = \hat{\boldsymbol{\tau}}, I_0 = \hat{I}_0, b = \hat{b}. \tag{6.7}$$

In this general case, the FIM is not block diagonal and the CRB is increased by having to estimate I_0 and b.

3. Auxiliary Data

One way to simplify the ML estimation of the parameters of interest and to avoid the increase in variance that results from having to estimate nuisance parameters is to acquire more data. Additional telescopes could be used to measure the guide-star brightness and sky background. Their collection apertures could be much larger than that of a single lenslet in a Shack–Hartmann sensor, and if scintillation effects are not important their integration time could be much longer. The resulting estimates of I_0

and b could have very low variance, so these parameters could be regarded as known.

If additional monitors are not practical, the data from all lenslets could be used to estimate I_0 and b. With J lenslets and the wavefront described by pure tilt, the complete data set is described by $2J+2$ parameters (two tilts per lenslets plus two global nuisance parameters), which is an improvement over the $4J$ we would have if two tilts and two nuisance parameters were to be estimated from the data associated with each lenslet. Even if scintillation does occur, it will not affect the diffuse sky background, so at least b can be treated as a global parameter.

B. Estimation of Global Wavefront Parameters

Above we considered the traditional operation of a Shack–Hartmann sensor in which the goal is assumed to be estimation of local tilts from data associated with individual lenslets. Once that is accomplished, the true goal of estimating global parameters in an expansion like Eq. (4.2) is often considered to be a separate problem.

This dichotomy is tenable in a Shack–Hartmann sensor only if radiation passing through one lenslet does not reach the detector pixels associated with an adjacent lenslet, but this condition is quite restrictive. Even if the detectors lie in the focal plane of the lenslet, the tails of the point-spread function from the lenslet of interest can overlap the pixels associated with an adjacent lenslet. Approaches to dealing with this problem and arriving at final ML estimates of global parameters are discussed below.

1. Likelihood Models with Overlap

Suppose we want to estimate local tilts using only the data from detector elements under a particular lenslet, even though light from other lenslets contributes to the data from those elements. We could simply ignore the problem and find ML estimates of the local tilts from an erroneous likelihood model. A rigorous mathematical treatment of the errors resulting from misspecified likelihood models is given by Halbert White,[50] who showed that there are many circumstances under which such quasi-ML estimators (QMLEs) have very useful properties. As with true ML estimators, the PDFs of QMLEs may asymptotically approach multivariate normals, though not necessarily with the inverse of the FIM as the covariance matrix, and they may be consistent estimators. White also gives several useful tests of the degree of misspecification of the likelihood model. No research has appeared on applying White's theories to wavefront sensing, so it is not yet clear what can be said about QMLEs of local tilts or when the likelihood specification is adequate.

Rather than ignoring the overlap problem, an alternative would be to treat the tilts in adjacent lenslets as nuisance parameters for the purpose of estimating the tilts over a given lenslet. Then the general theory developed in Subsection 5.B would be applicable and a multivariate normal model, like Eq. (5.7) but with the 2D vector $\boldsymbol{\tau}$ in place of $\boldsymbol{\alpha}$, would result after marginalization.

Finally, we could consider inserting physical dividers between the lenslets to prevent the overlap, ensuring that the local likelihood model was valid. An immediate consequence would be that assumption (6.2) would not hold and hence it would be necessary to estimate the guide-star brightness (or measure it independently) along with the local tilt.

2. Maximum-Likelihood Estimation of Mirror-Mode Coefficients

There are several possible ways of getting ML estimates of the vector of mirror-mode coefficients $\boldsymbol{\alpha}$, depending on what we use as the initial data.

If we have valid ML estimates of local tilts, we may be able to get ML estimates of $\boldsymbol{\alpha}$ by use of the ML invariance principle (2.15), at least when J (the number of lenslets) and N (the number of mirror actuators) are both large. Details of this approach and conditions for its validity are given in Appendix C, but the conditions are difficult to meet in practice.

Alternatively, if we have any estimates at all of local tilts, even centroid estimates, we can use them as data from which to estimate $\boldsymbol{\alpha}$ so long as we can construct the relevant likelihood model. If we denote the estimates as $\hat{\boldsymbol{\tau}}$, the likelihood we need is $\mathrm{pr}(\hat{\boldsymbol{\tau}}|\boldsymbol{\alpha})$. As we show in Appendix D, however, finding the relevant likelihood can be complicated, and without an accurate likelihood, neither ML nor MAP estimation of $\boldsymbol{\alpha}$ can be considered optimal in any sense.

A better approach is to start with the raw data $\mathbf{g}$ (the detector outputs $\{g_m\}$ for all m, not just the ones associated with a single lenslet). The likelihood function in that case is $\mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha})$, which is just what we have been discussing throughout this paper. Any of the likelihood models from Section 5 can be used.

C. Simulation Results

To illustrate the theory developed in this paper, we performed several simulation studies of a Shack–Hartmann sensor.

In the first study, designed to test the ability of the ML method to reduce nonlinearity in a Shack–Hartmann sensor, only a single lenslet was considered, and a 2×2 array of photodetectors (often called a quad cell) was placed in its focal plane. The irradiance for a given tilt, $s(\mathbf{r}-z_0\boldsymbol{\tau})$ in Eq. (6.1) was assumed to be a 2D Gaussian function, and the mean response functions, $f_m(\boldsymbol{\tau})$, $m=1,\ldots,4$, were found by performing the integral in Eq. (6.1) numerically; the results are shown in Fig. 3.

These response functions were then used to generate pure Poisson data for an 8×8 array of tilts. For each position in the array, 200 realizations of a 4D Poisson random vector (one component for each detector in the quad cell) were generated. These data were used in both a standard centroid estimator (see Appendix D) and a simple ML estimator based on the Poisson statistics. There were no nuisance parameters, and the log-likelihood was given by Eq. (6.3) with I_0 assumed known. The maximization of the likelihood was performed by a Nelder–Mead algorithm implemented in the Matlab function fminsearch. Each of the resulting estimates was plotted as a point in a 2D image, one image for the centroid estimates and one for ML. These images, shown in Fig. 4, are thus approximations to the PDFs of the tilt estimates when the true values are delta functions on an 8×8 array of points.

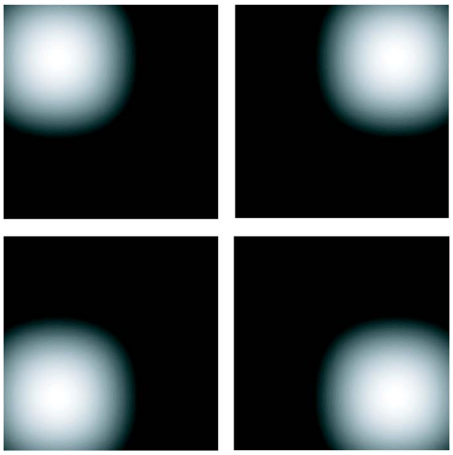

Fig. 3. (Color online) Display of the response functions $f_m(\tau)$ used in simulation of a Shack–Hartmann sensor with a single lenslet and a 2×2 array of photodetectors. Each plot represents the mean response of one photodetector as a function of the x and y components of the wavefront tilt.

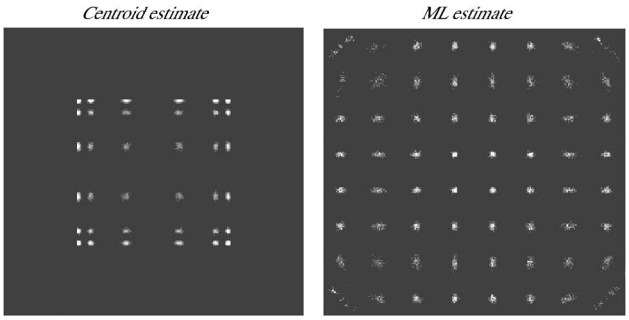

Fig. 4. Left, centroid estimates of an 8×8 array of tilts from Poisson data in a quad-cell Shack–Hartmann sensor; right, ML estimates from the same data.

With the centroid estimator, only a 6×6 array of points is seen on the left in Fig. 4; the outermost points overlap with their neighbors, and information about these larger tilts is irretrievably lost. This problem cannot be eliminated by any form of nonlinearity correction; no transformation of the left image in Fig. 4 can remove the complete overlap of the outermost points with their neighbors. With the ML estimator, on the other hand, the nonlinearity is almost completely eliminated (the estimator is nearly unbiased), and the dynamic range of the quad-cell sensor is approximately doubled. Both estimators are nearly unbiased and efficient for a point in the center of the array.

A more extensive comparison of ML and centroid estimations of tilts, taking account of nuisance parameters and null functions and exploring a much wider range of noise characteristics and photodetector arrays, will be published separately.

A second simulation study considered estimation of wavefront parameters directly from photodetector outputs without an intermediate estimation of tilts. A wavefront aberration was simulated using the 12 Zernike polynomials between the 2nd and the 4th radial order with positive coefficients that followed Noll's[51] mean-square residual error distribution for $D/r_0=16$ (the total wavefront rms was 3.28 rad). A pixellated (CCD) image of the spot pattern of the wavefront aberration in a Shack–Hartmann sensor was simulated on a computer using the discrete Fourier transform (DFT) implementation of the Fresnel diffraction formula. The simulated detector had 128×128 square pixels, and the Shack–Hartmann sensor had 16 square lenslets across the diameter of the full pupil (8×8 pixels on the detector for each lenslet). The focal length of the lenslets was set to approximately 50 times the lateral size of each lenslet.

Fifty realizations of pure Poisson deviates of the CCD image were generated for each of six different light levels: $10^{-1/2}$, 1, $10^{-1/2}$, 10, $10^{3/2}$, and 100 photons/lenslet. The coefficients of the 12 Zernike polynomials included in the

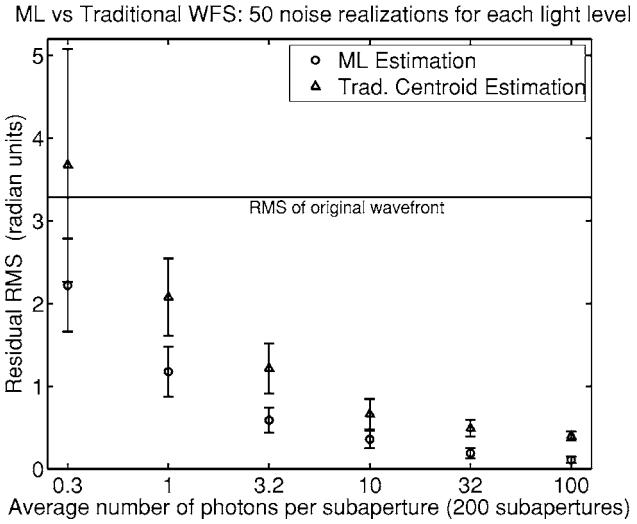

Fig. 5. Comparison of traditional LS estimation of wavefront coefficients from centroid data versus direct ML estimation from photodetector outputs. Parameters used in the simulation include: $\lambda=680$ nm; pupil diameter$=24\ \mu\text{m}\times128=3072\ \mu$m; lenslet size$=192\ \mu$m; CCD pixel size$=24\ \mu$m; and focal length$=9.9$ mm. The wavefront was sampled at 1726 points across the pupil diameter, and 322 rows and columns of zeros were used to pad the wavefront function to a 2048×2048 array before computing the FFT. The markers represent the mean, and the error bars represent the standard deviation of the residual wavefront rms of the 50 estimations for each light level.

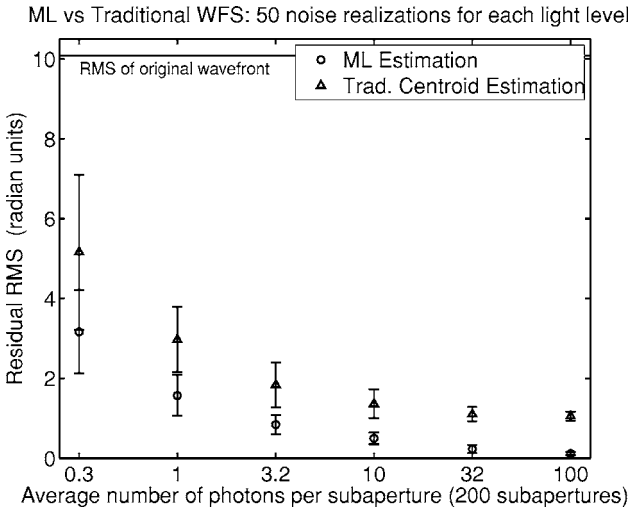

Fig. 6. Same as Fig. 5 except that global tip and tilt were not removed from the simulated wavefront and were also included in the coefficients to estimate.

wavefront were estimated from the same data by using both ML and traditional centroiding with LS reconstruction. The results of the simulations are shown in Fig. 5. In another study, the aberrated wavefront also included global tip and tilt, for a total of 14 unknown coefficients; the results in that case are shown in Fig. 6.

As seen from the figures, direct ML estimation can offer up to a fourfold advantage in residual wavefront error (ninefold if the global tip and tilt terms are not corrected separately), suggesting that there is indeed a significant loss of information in the tilt-estimation step (the preprocessing stage in Fig. 1). Such a loss is not surprising since tilt estimation in this case reduces an 8×8 array of photodetector outputs to just two centroids.

It is also noteworthy that a significant reduction in wavefront error can be achieved with an average of 0.32 photons/subaperture, or 0.005 photons/detector element. Of course this level of performance would not be obtained if sky background or readout noise were considered, but it is possible that ML methods would have even larger advantages over traditional methods in these cases because of more accurate statistical modeling. A detailed study of these issues is in progress.

7. COMPUTATIONAL METHODS

Astronomical WFSs must respond on a time scale of 10–100 ms, depending on wavelength and wind speed, and any computations performed by the sensor must be at least this fast. Since ML estimation usually uses an iterative search for the maximum, it might seem difficult to meet this requirement, but we can draw on methods developed for the closely analogous problem of ML position estimation in scintillation cameras for gamma-ray imaging. In that application, the computation must be carried out in a few microseconds rather than milliseconds, but hardware and software approaches that meet this goal have been demonstrated. In this section we summarize these approaches and then discuss how they can be applied to wavefront sensing.

A. Computational Approaches from Gamma-Ray Imaging

In a scintillation camera, a gamma ray interacts in a scintillation crystal such as sodium iodide and produces a flash of light that illuminates an array of PMTs. The objective is to determine the coordinates of the interaction event and the strength of the light flash, which is proportional to the gamma-ray energy. Since the estimate must be obtained for each gamma-ray photon, and the photons arrive randomly at mean rates that can exceed 10^5 events/s, it is desirable to carry out the estimation in 1 to 2 μs.

If the scintillation crystal is relatively thin, it suffices to estimate the lateral coordinates (x,y) of the scintillation event, but at high gamma-ray energies a thicker crystal must be used, and the z coordinate (normal to the entrance face of the crystal) also influences the data. Depending on the application, the z coordinate, referred to as the depth of interaction, can be regarded as a nuisance parameter or as another parameter to estimate. If the variables to be estimated are x, y and the brightness of the flash I_0, then the estimation problem in a scintillation camera is equivalent to estimating the two components of tilt and the guide-star brightness in wavefront sensing.

In some problems two gamma rays can be absorbed simultaneously in the scintillation crystal, either because the radioisotope emits two photons in a rapid cascade or because of Compton scatter in the crystal. In these cases the number of parameters to estimate can be as large as eight (three spatial coordinates and energy for each of two photons). Alternatively, the properties of the secondary photon can be treated as additional nuisance parameters.

For the scintillation cameras developed at the Center for Gamma-Ray Imaging of the University of Arizona, the data dimension M is either 4 (a 2×2 array of photomultipliers), 9 (a 3×3 array), or 64 (an 8×8 array). Thus the goal of the processing is to estimate a set of 2–8 parameters from a set of 4–64 measurements in about 2 μs.

The statistical models used with scintillation cameras are remarkably similar to those considered in this paper.[52] In most cases the log-likelihoods have the structure

$$\ln \mathrm{pr}(\mathbf{g}|\boldsymbol{\Theta}) = \sum_{m=1}^{M} \ln \mathrm{pr}[g_m|\bar{g}_m(\boldsymbol{\Theta})], \tag{7.1}$$

where $\boldsymbol{\Theta}$ is the set of parameters to be estimated. In this paper the only log-likelihood not in the form of Eq. (7.1) is Eq. (5.25), where a correlated multivariate normal was obtained by marginalizing over intrinsic nuisance parameters. Similarly, in a scintillation camera, a multivariate normal can be used to describe the likelihood that results from marginalizing over the depth of interaction.

When the log-likelihoods have the form of Eq. (7.1), their dependence on $\boldsymbol{\Theta}$ is determined by the set of means $\{\bar{g}_m(\boldsymbol{\Theta})\}$, which we refer to as mean detector response functions or MDRFs.[27,28] The MDRFs can either be measured directly with a collimated source of gamma rays or be simulated by an optical transport code that models the camera. Once they are known, they can be stored as look-up tables, even when the dimension of $\boldsymbol{\Theta}$ is as large as 8. For $N=2$, when the problem is just to estimate the (x,y) coordinates of each scintillation event, then each $\bar{g}_m(\boldsymbol{\Theta})$ can be stored as a $K_x\times K_x$ image, where K_x is the number of discretization steps in x or y ($K_x=128$ or 256, say). Even with 64 PMTs, therefore, the storage requirements are modest. If we add the depth of interaction z as a parameter to estimate, the necessary storage increases by a factor of K_z, the number of steps in z, but this is typically only 10 or so. Adding the photon energy to the list of parameters requires no additional storage since the MDRF factorizes in the same way as in Eq. (5.10). Estimating the coordinates of two simultaneous events increases N to 8 but does not increase the storage required for the MDRFs, since the total light incident on any PMT from the two events is just the sum of the contributions from the individual event, to a good approximation.

With stored MDRFs, evaluation of the log-likelihood at any $\boldsymbol{\Theta}$ can be accomplished rapidly by looking up the value of $\bar{g}_m(\boldsymbol{\Theta})$ for each m, using a second look-up table to find each $\ln \mathrm{pr}[g_m|\bar{g}_m(\boldsymbol{\Theta})]$, and adding the results. The second look-up table has $K_m\times K_{\bar{g}}$ entries, where K_m is determined by the analog-to-digital (A/D) converter used to

digitize the photomultiplier signals and $K_{\bar{g}}$ is related to the resolution used for $\bar{g}_m(\boldsymbol{\Theta})$; the dimension of $\boldsymbol{\Theta}$ is irrelevant in this table.

Since it is not so easy to compute and store derivatives of the MDRFs, search algorithms for finding the ML estimates in scintillation cameras have concentrated on methods that require the value of the log-likelihood but not gradients or Hessians. For searching over just the x and y coordinates, it is feasible to choose a reasonable starting point, say the coordinates of the PMT that gets the largest signal, and then do an exhaustive search over a subset of x and y in this vicinity.

Exhaustive search fails when additional parameters are to be estimated, and in those cases useful search algorithms include iterative coordinate descent, variations on the Nelder–Mead simplex, and multigrid algorithms. Iterative coordinate descent performs a sequence of 1D searches on each of the N individual components of $\boldsymbol{\Theta}$ in turn, while simplex methods compute the log-likelihoods on a set of $N+1$ points in the N-dimensional parameter space at each iteration and use some rules for modifying the coordinates of the points in order to go to the next iteration. Multigrid techniques are similar to simplex methods in that the log-likelihood is computed on a set of points at each iteration, but the points are regularly spaced in parameter space; a coarse spacing is used initially and is then reduced as the iteration proceeds. Conjugate gradient searches, as suggested by Cannon[21] for global wavefront estimation, are very effective when gradients can be calculated analytically. All of these methods work well when the function being searched is smooth and unimodal, as is usually the case with log-likelihoods for scintillation cameras.

Furenlid[31] and Hesterman[53] have recently implemented a multigrid method for scintillation cameras. In initial experiments, a 4×4 grid of points was used in a two-dimensional parameter space, and the grid spacing was halved at each iteration. The algorithm converged in six iterations to exactly the same estimates as those found by an exhaustive search. The calculation requires 16 μs in C on a single Macintosh G5 computer, but Furenlid has shown that it can be converted to a pipeline process in a field-programmable gate array (FPGA). In that case all likelihood calculations are done in parallel, and J iterations of the algorithm require just J clock cycles, where each clock cycle is a few nanoseconds with modern FPGAs. There should be no difficulty in principle in using a similar pipeline architecture with a simplex search.

Finally, we mention that for the special case of estimation from four measurements, as with a 2×2 array of photodetectors, the entire search process can be performed offline and stored in a look-up table for all possible combinations of the four signals. If K_m A/D levels are used for each measurement, then there are 4^{K_m} locations in the table, and the final ML estimate of up to four parameters can be stored at each location. A useful practical trick is to take the square root of the measurements before coarse discretization in order to make the variance approximately constant, and with this measure it is found that 6-bit quantization ($K_m=64$) suffices, so the look-up table is easily stored in memory. No real-time search is needed, and the estimate is available in the time required to do a single memory access. This method has been used routinely for two decades with four-PMT scintillation cameras at the University of Arizona.[27,28]

B. Methods for Maximum-Likelihood Estimation in Wavefront Sensing

The methods discussed above for scintillation cameras are immediately applicable to estimation of tilts over one subaperture of a Shack–Hartmann sensor, even with one or two nuisance parameters. The multigrid method with an FPGA devoted to each subaperture will give the estimate in less than a microsecond for any realistic number of detectors per subaperture, and data from multiple subapertures can readily be multiplexed through a single FPGA. Even when the multigrid method is implemented on a single processor, it appears that it will allow estimation of all subaperture tilts in less than a millisecond. Moreover, if a Shack–Hartmann sensor with nanosecond response should ever be required, it can be achieved by using 2×2 arrays of fast detectors at each subaperture and look-up tables for the final ML estimates of subaperture tilts.

The computational difficulties in ML estimation increase with the number of parameters being estimated and the number of independent measurements, and it is not so obvious that the speed requirements for astronomical wavefront sensing can be met if we choose to estimate a large number of modal coefficients $\{\alpha_n,\ n=1,\ldots N\}$ from the entire set of detector measurements directly. If these coefficients specify the possible configurations of a deformable mirror, then N is the number of actuators, which ranges from 20 to 40 in laboratory systems to hundreds or even thousands in large telescopes.

The dimension of the data vector is also a concern. The number of independent measurements does not exceed the number of pixels in the detector array in the wavefront sensor, but in many cases it can be much less. With a Shack–Hartmann or any other sensor that divides the wavefront into subapertures, the local parameters associated with one subaperture (e.g., local tilts and/or curvatures) can be estimated from the data associated with that subaperture. Moreover, many of the data values will be near zero in practice and can be omitted from the data vector. For example, a diffraction-limited spot in a Shack–Hartmann sensor will illuminate a fraction $\sim(\lambda f_l)^2/D_l^4$ of the detector pixels, where D_l is the diameter of the lenslet and f_l is its focal length; other pixels can be set to zero by thresholding. Similarly, if the readout noise is low enough that a single photon can be detected, the number of nonzero measurements after thresholding does not exceed the number of detected photons.

The dimension of the data vector used for an estimation problem can be also reduced by computing functions of the raw data called sufficient statistics. By definition, a set of sufficient statistics contains the same information about the estimation problem as the raw data does, but if the dimension of the set is much less than the number of original measurements, a considerable computational saving can be achieved. There is some current activity in finding sufficient statistics for position estimation in scintillation cameras,[52] and these methods are potentially useful in wavefront sensing as well.

The complexity of the search algorithm depends on the dimensions of both the data and the parameter space. To illustrate the point, consider the Poisson model [Eq. (5.9)], where the only nuisance parameter is a global guide-star brightness I_0; this model is valid if N is large and there is no atmospheric scintillation. Under these same assumptions, the total light reaching the detector plane is independent of the wavefront parameters α, and if the detectors are identical and there are no gaps between them, we can write

$$\sum_{m=1}^{M} f_m(\boldsymbol{\alpha}) = f_{tot} = \text{constant}. \tag{7.2}$$

This assumption for modal estimation is more defensible than its counterpart for local tilt estimation, [Eq. (6.2)], since we do not need to worry about light that misses a subset of the detectors or overlap of light from different subapertures; Eq. (7.2) is simply a statement of conservation of energy. With this model, the ML estimate of I_0 is just $\hat{I}_0 = N_{tot}/f_{tot}$ [cf. Eq. (6.5)], and the ML estimate of $\boldsymbol{\alpha}$ must satisfy [cf. Eq. (6.3)]

$$\sum_{m=1}^{M} g_m \ln[f_m(\boldsymbol{\alpha})] = \text{maximum}. \tag{7.3}$$

The functions $\{f_m(\boldsymbol{\alpha})\}$ are the counterparts of the MDRFs for scintillation cameras, but there are more of them and each is a function in a higher-dimensional space. Precomputing and storing them is difficult, and the feasibility of ML estimation of the modal parameters depends on being able to compute the $f_m(\boldsymbol{\alpha})$ rapidly.

There are several factors that simplify the problem. First, numerical studies (to be published separately) show that the log-likelihood for the modal parameters is smooth and slowly varying, especially at low light level. Thus it suffices to compute $f_m(\boldsymbol{\alpha})$ on a sparse grid in parameter space and use, say, spline interpolation to find it at intermediate points.

Second, in almost all applications the parameters will change slowly from frame to frame of the wavefront-sensor data, so an estimate found on one frame will be an excellent starting point for the next frame. Moreover, in a closed-loop system with good correction, we need to search only in the vicinity of the origin of parameter space, where all $\alpha_n = 0$.

Third, the problem is amenable to parallel computation in several possible ways. In a simplex or multigrid algorithm, for example, different processors can be assigned to different points in parameter space. In an N-dimensional estimation problem, a simplex requires computing the log-likelihoods at $N+1$ values of $\boldsymbol{\alpha}$, which can be performed with $N+1$ processors. If a full diffraction-theory model is used for the computation, the use of dedicated fast Fourier transform (FFT) chips in each processor might be advantageous.

A less obvious way to parallelize the problem is to divide the data space into subsets, perhaps corresponding to subapertures even if the goal is not to estimate local tilts. The advantage of this division is that the wavefront in the local region is described by a small set of parameters such as the local tilts and curvatures, and these local parameters are easily computed as linear combinations of the components of interest $\{\alpha_n\}$. With this simplification we are back to efficient calculations or even look-up tables to find the values $\ln[f_m(\boldsymbol{\alpha})]$ for each m in the data subset, and the overall log-likelihood is found by collecting the results from individual processors and summing as in Eq. (7.3). Again, simplex or multigrid methods can be used for efficient search without computing derivatives.

8. SUMMARY AND CONCLUSIONS

Maximum-likelihood estimation offers several theoretical advantages in general. An ML estimate is efficient if an efficient estimator exists, and it is asymptotically unbiased, efficient, and consistent as more data are acquired in any case. Compared with other computational methods in wavefront sensing, ML can reduce the bias and variance of the estimates of tilts, modal coefficients, or any other wavefront parameters, basically by taking advantage of the knowledge of the data statistics and using a more accurate model of the deterministic properties of the sensor. Unlike MAP or other Bayesian estimates, ML estimates do not incorporate any prior knowledge of the parameters to be estimated, but accurate likelihood models are essential to good MAP estimation also.

It is relatively straightforward to write down conditional PDFs for the data produced by the detectors in a wavefront sensor, but these PDFs are not the likelihoods needed for ML (or MAP) estimation of wavefront parameters for two reasons. First, not all parameters associated with the wavefront influence the data significantly; the ones that do not are called null functions. Second, there may be parameters that do influence the data but that we are not interested in estimating; they are called nuisance parameters. This paper has been concerned largely with the effect of null functions and nuisance parameters in wavefront sensing.

The basic stochastic models considered here included Poisson noise from the photoelectron statistics, Gaussian noise from the electronics, and a mixture of the two. Excess noise from detectors with internal gain was not considered explicitly, but most of the theory is easily adapted to that case. As in all ML problems, the parameters to be estimated were not considered to be random, but nuisance parameters were, and the final likelihoods of interest were obtained by marginalizing with respect to some prior distribution on the nuisance parameters. General expressions for both log-likelihoods and FIMs were derived on this basis. The theory was illustrated by discussing the estimation of local tilts and modal parameters from Shack–Hartmann data.

Computational issues associated with both the Shack–Hartmann subaperture problem and the more general problem of estimating coefficients in a modal expansion of the wavefront were discussed. For the subaperture case it was seen that ML estimation in microseconds or even nanoseconds is feasible, and several approaches that should lead to millisecond computation of modal coefficients were outlined. Work on the latter problem is actively underway and will be reported at a later date.

APPENDIX A: FISHER INFORMATION MATRIX FOR COMBINED POISSON AND GAUSSIAN NOISE

In this appendix we derive the FIM with both Poisson and Gaussian noise. The basic statistical model is the Poisson–Gaussian mixture developed by Snyder *et al.*[41] The starting point for this appendix is Eq. (3.16), which for a single detector element can be written without the subscripts as

$$\mathrm{pr}(g|\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{k=0}^{\infty} \exp\left[-\frac{(g-Rk)^2}{2\sigma^2} \right] \frac{[\bar{k}(\boldsymbol{\theta})]^k}{k!} \exp[-\bar{k}(\boldsymbol{\theta})]. \tag{A1}$$

The FIM is the covariance matrix of the score vector, defined as the gradient of the log-likelihood with respect to the parameters being estimated. For the PDF of Eq. (A1), the nth component of the score is given by

$$\begin{aligned} \frac{\partial}{\partial\theta_n} \ln \mathrm{pr}(g|\boldsymbol{\theta}) &= \frac{1}{\mathrm{pr}(g|\boldsymbol{\theta})} \frac{\partial}{\partial\theta_n} \mathrm{pr}(g|\boldsymbol{\theta}) = \frac{1}{\mathrm{pr}(g|\boldsymbol{\theta})} \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{k=0}^{\infty} \exp\left[-\frac{(g-Rk)^2}{2\sigma^2} \right] \frac{1}{k!} \frac{\partial}{\partial\theta_n} \{[\bar{k}(\boldsymbol{\theta})]^k \exp[-\bar{k}(\boldsymbol{\theta})]\} \\ &= \frac{1}{\mathrm{pr}(g|\boldsymbol{\theta})} \frac{1}{\sqrt{2\pi\sigma^2}} \sum_{k=0}^{\infty} \exp\left[-\frac{(g-Rk)^2}{2\sigma^2} \right] \frac{[\bar{k}(\boldsymbol{\theta})^k]}{k!} \exp[-\bar{k}(\boldsymbol{\theta})] \left(\frac{1}{\bar{k}(\boldsymbol{\theta})} - 1 \right) \frac{\partial \bar{k}(\boldsymbol{\theta})}{\partial\theta_n}. \end{aligned} \tag{A2}$$

A change of variables $k' = k-1$ and some algebra yields

$$\frac{\partial}{\partial\theta_n} \ln \mathrm{pr}(g|\boldsymbol{\theta}) = \left[\frac{\sum_{k'=0}^{\infty} \exp[-(1/2\sigma^2)(g-R-Rk')^2] \exp[-\bar{k}(\boldsymbol{\theta})][\bar{k}(\boldsymbol{\theta})]^{k'}/k'!}{\sum_{k=0}^{\infty} \exp[-(1/2\sigma^2)(g-Rk)^2] \exp[-\bar{k}(\boldsymbol{\theta})][\bar{k}(\boldsymbol{\theta})]^k/k!} - 1 \right] \frac{\partial \bar{k}(\boldsymbol{\theta})}{\partial\theta_n}. \tag{A3}$$

The difference between numerator and denominator is in the shift of the Gaussian factor.

A more explicit notation may clarify the result; if we let $\mathrm{pr}(g|\boldsymbol{\theta})$ be denoted by $\mathrm{pr}_{g|\boldsymbol{\theta}}(g)$ to indicate a specific function of g, then Eq. (A3) becomes

$$\frac{\partial}{\partial\theta_n} \ln \mathrm{pr}_{g|\boldsymbol{\theta}}(g) = \left[\frac{\mathrm{pr}_{g|\boldsymbol{\theta}}(g-R)}{\mathrm{pr}_{g|\boldsymbol{\theta}}(g)} - 1 \right] \frac{\partial \bar{k}(\boldsymbol{\theta})}{\partial\theta_n} \tag{A4}$$

or

$$\frac{\partial}{\partial\theta_n} \mathrm{pr}_{g|\boldsymbol{\theta}}(g) = [\mathrm{pr}_{g|\boldsymbol{\theta}}(g-R) - \mathrm{pr}_{g|\boldsymbol{\theta}}(g)] \frac{\partial \bar{k}(\boldsymbol{\theta})}{\partial\theta_n}. \tag{A5}$$

Since $\mathrm{pr}_{g|\boldsymbol{\theta}}(g)$ is, for example, the PDF depicted in Fig. 2(a), and $\mathrm{pr}_{g|\boldsymbol{\theta}}(g-R)$ is the same function shifted to the right by an amount R (i.e., shifted over one peak in Fig. 2(a), Eq. (A5) looks like the chain rule of differentiation with one derivative replaced by a finite difference, but in fact the result is exact.

Elements of the FIM (for a single detector) are given by

$$\begin{aligned} F_{nn'} &= \left\langle \left[\frac{\partial}{\partial\theta_n} \ln \mathrm{pr}_{g|\boldsymbol{\theta}}(g) \right] \left[\frac{\partial}{\partial\theta_{n'}} \ln \mathrm{pr}_{g|\boldsymbol{\theta}}(g) \right] \right\rangle_{g|\boldsymbol{\theta}} \\ &= \left\langle \left[\frac{\mathrm{pr}_{g|\boldsymbol{\theta}}(g-R)}{\mathrm{pr}_{g|\boldsymbol{\theta}}(g)} - 1 \right]^2 \right\rangle_{g|\boldsymbol{\theta}} \frac{\partial \bar{k}(\boldsymbol{\theta})}{\partial\theta_n} \frac{\partial \bar{k}(\boldsymbol{\theta})}{\partial\theta_{n'}}. \end{aligned} \tag{A6}$$

The expectation can be written in detail as

$$\begin{aligned} &\left\langle \left[\frac{\mathrm{pr}_{g|\boldsymbol{\theta}}(g-R)}{\mathrm{pr}_{g|\boldsymbol{\theta}}(g)} - 1 \right]^2 \right\rangle_{g|\boldsymbol{\theta}} \\ &\quad = \int_{-\infty}^{\infty} \mathrm{d}g\, \mathrm{pr}_{g|\boldsymbol{\theta}}(g) \left[\frac{\mathrm{pr}_{g|\boldsymbol{\theta}}(g-R)}{\mathrm{pr}_{g|\boldsymbol{\theta}}(g)} - 1 \right]^2 \\ &\quad = \int_{-\infty}^{\infty} \mathrm{d}g \frac{[\mathrm{pr}_{g|\boldsymbol{\theta}}(g-R)]^2}{\mathrm{pr}_{g|\boldsymbol{\theta}}(g)} - 1, \end{aligned} \tag{A7}$$

where the normalization of PDFs has been used to get the second line.

Thus the FIM for one detector element is given by

$$F_{jk} = \left[\int_{-\infty}^{\infty} \mathrm{d}g \frac{[\mathrm{pr}_{g|\boldsymbol{\theta}}(g-R)]^2}{\mathrm{pr}_{g|\boldsymbol{\theta}}(g)} - 1 \right] \frac{\partial \bar{k}(\boldsymbol{\theta})}{\partial\theta_j} \frac{\partial \bar{k}(\boldsymbol{\theta})}{\partial\theta_k}. \tag{A8}$$

This expression is exact and numerically tractable since the integral is one dimensional.

The reader versed in statistical decision theory will recognize $\mathrm{pr}_{g|\boldsymbol{\theta}}(g-R)/\mathrm{pr}_{g|\boldsymbol{\theta}}(g)$ as a likelihood ratio Λ. The likelihood ratio is the ideal test statistic for deciding between two hypotheses, in this case the null hypothesis H_0 that g is drawn from the unshifted density $\mathrm{pr}_{g|\boldsymbol{\theta}}(g)$ and the alternative hypothesis H_1 that g is drawn from $\mathrm{pr}_{g|\boldsymbol{\theta}}(g-R)$. With that interpretation, the integral in Eq. (A8) is the expectation of Λ under H_1, a quantity that is closely related to performance on discrimination tasks,[6] and Eq. (A8) establishes a relationship between that discrimination task and the estimation task that is the subject of this paper.

The factor in square brackets in Eq. (A8) can be evaluated in several limits. For pure Poisson noise ($\sigma^2 \to 0$), it is

$1/\bar{k}(\boldsymbol{\theta})$. Pure Gaussian noise corresponds to the limit $\bar{k}(\boldsymbol{\theta})\to\infty$ and $R\to 0$ in such a way that $R\bar{k}(\boldsymbol{\theta})$ remains constant, and in that limit the factor tends to R^2/σ^2. Numerical studies show that a useful approximate form in all cases (even when the PDF is highly non-Gaussian) is

$$F_{jk} \approx \frac{R^2}{\sigma^2 + R^2\bar{k}(\boldsymbol{\theta})} \frac{\partial \bar{k}(\boldsymbol{\theta})}{\partial \theta_j} \frac{\partial \bar{k}(\boldsymbol{\theta})}{\partial \theta_k}. \tag{A9}$$

If there are M detectors but the measurements are statistically independent, as we assumed in Subsection 3.D, then the final expression for the FIM [Eq. (3.17)], is obtained by reinstating the subscripts on g_m and $\bar{k}_m(\boldsymbol{\theta})$ and then summing over m.

APPENDIX B: MARGINALIZING OVER NUISANCE PARAMETERS

In this appendix we fill in some details needed in Subsection 5.B regarding marginalizing intrinsic nuisance parameters. Extrinsic nuisance parameters are not considered here, so the $P\times 1$ vector of all parameters that influence the data can be written as $\boldsymbol{\theta}=(\boldsymbol{\alpha},\boldsymbol{\beta})^t$, where $\boldsymbol{\alpha}$ is $N\times 1$, $\boldsymbol{\beta}$ is $K\times 1$, and $N+K=P$.

It is assumed that the prior PDF describing $\boldsymbol{\beta}$ is a multivariate normal of the form

$$\mathrm{pr}(\boldsymbol{\theta}) = \mathrm{pr}(\boldsymbol{\alpha},\boldsymbol{\beta}) = \mathcal{N}_{\boldsymbol{\theta}} \exp\Big[-\tfrac{1}{2}(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})^t \mathbf{K}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})\Big], \tag{B1}$$

where the covariance matrix can be written in the partitioned form

$$\mathbf{K}_{\boldsymbol{\theta}} = \begin{bmatrix} \mathbf{K}_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & \mathbf{K}_{\boldsymbol{\alpha}\boldsymbol{\beta}} \\ \mathbf{K}_{\boldsymbol{\alpha}\boldsymbol{\beta}}^t & \mathbf{K}_{\boldsymbol{\beta}\boldsymbol{\beta}} \end{bmatrix}, \tag{B2}$$

and $\mathcal{N}_{\boldsymbol{\theta}}=[(2\pi)^N \det(\mathbf{K}_{\boldsymbol{\theta}})]^{-1/2}$ is the normalizing constant.

Some well-known results from multivariate statistics[54,55] show that the marginal density needed in Eq. (5.2) has the form

$$\mathrm{pr}(\boldsymbol{\beta}|\boldsymbol{\alpha}) = \mathcal{N}_{\boldsymbol{\beta}|\boldsymbol{\alpha}} \exp\Big[-\tfrac{1}{2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^t \mathbf{K}_{\boldsymbol{\beta}|\boldsymbol{\alpha}}^{-1}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\Big], \tag{B3}$$

where

$$\tilde{\boldsymbol{\beta}} = \bar{\boldsymbol{\beta}} + \mathbf{K}_{\boldsymbol{\beta}\boldsymbol{\alpha}}\mathbf{K}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1}(\boldsymbol{\alpha}-\bar{\boldsymbol{\alpha}}), \tag{B4}$$

$$\mathbf{K}_{\boldsymbol{\beta}|\alpha} = \mathbf{K}_{\boldsymbol{\beta}\boldsymbol{\beta}} - \mathbf{K}_{\boldsymbol{\beta}\boldsymbol{\alpha}}\mathbf{K}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1}\mathbf{K}_{\boldsymbol{\alpha}\boldsymbol{\beta}}. \tag{B5}$$

The matrix $\mathbf{K}_{\boldsymbol{\beta}|\boldsymbol{\alpha}}$, which arises from taking the inverse of a partitioned matrix, is known as the Schur complement of $\mathbf{K}_{\boldsymbol{\alpha}\boldsymbol{\alpha}}$. The results in Eqs. (B3)–(B5) are specialized to a wavefront sensor used in a closed-loop AO system discussed in Subsection 5.B.

We also need to evaluate the integral in Eq. (5.6) when the intrinsic nuisance parameters make a small perturbation to the mean data, in which case we can expand the mean data as

$$\bar{g}_m(\boldsymbol{\alpha},\boldsymbol{\beta}^{ext},\boldsymbol{\beta}^{int}) \approx \bar{g}_m(\boldsymbol{\alpha},\boldsymbol{\beta}^{ext},\mathbf{0}) + \sum_{k=1}^{K} A_{mk}\beta_k^{int}$$

$$\text{where} \quad A_{mk} = \left.\frac{\partial \bar{g}_m(\boldsymbol{\alpha},\boldsymbol{\beta}^{ext},\boldsymbol{\beta}^{int})}{\partial \beta_k^{int}}\right|_{\boldsymbol{\beta}^{int}=\mathbf{0}}. \tag{B6}$$

For notational simplicity we let $\bar{\mathbf{g}}(\boldsymbol{\alpha},\boldsymbol{\beta}^{ext},\boldsymbol{\beta}^{int})=\bar{\mathbf{g}}$ and $\bar{\mathbf{g}}(\boldsymbol{\alpha},\boldsymbol{\beta}^{ext},\mathbf{0})=\bar{\mathbf{g}}_0$, so Eq. (B11) reads

$$\bar{\mathbf{g}} = \bar{\mathbf{g}}_0 + \mathbf{A}\boldsymbol{\beta}^{int}. \tag{B7}$$

Then the integral in Eq. (5.6) can be written as

$$\mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha},\boldsymbol{\beta}^{ext}) \approx \mathcal{N}' \int \mathrm{d}^K\beta^{int} \times \exp\left[-\frac{1}{2\sigma^2}\|\mathbf{g}-\bar{g}_0-\mathbf{A}\boldsymbol{\beta}^{int}\|^2\right] \times \exp\Big[-\tfrac{1}{2}(\boldsymbol{\beta}^{int})^t\mathbf{C}^{-1}(\boldsymbol{\beta}^{int})\Big], \tag{B8}$$

We can perform the integral by representing each probability density in terms of its characteristic function. The PDF of an M-dimensional multivariate normal vector $\mathbf{x}$ of mean $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{K}$ can be written as

$$\mathrm{pr}(\mathbf{x}) = [(2\pi)^M \det(\mathbf{K})]^{-1/2}\exp\Big[-\tfrac{1}{2}(\mathbf{x}-\bar{\mathbf{x}})^t\mathbf{K}^{-1}(\mathbf{x}-\bar{\mathbf{x}})\Big] = \int_{\infty} \mathrm{d}^M\xi \exp[2\pi i\boldsymbol{\xi}^t(\mathbf{x}-\bar{\mathbf{x}})]\exp(-2\pi^2\boldsymbol{\xi}^t\mathbf{K}\boldsymbol{\xi}). \tag{B9}$$

Expanding both densities in Eq. (B8) this way yields

$$\mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha},\boldsymbol{\beta}^{ext}) = \int_{\infty}\mathrm{d}^K\beta^{int}\int_{\infty}\mathrm{d}^M\xi\int_{\infty}\mathrm{d}^K\eta\exp(-2\pi^2\sigma^2\|\boldsymbol{\xi}\|^2) \times \exp[-2\pi i\boldsymbol{\xi}^t(\bar{\mathbf{g}}_0+\mathbf{A}\boldsymbol{\beta}^{int})]\exp(2\pi i\boldsymbol{\xi}^t\mathbf{g}) \times \exp(-2\pi^2\boldsymbol{\eta}^t\mathbf{C}\boldsymbol{\eta})\exp(2\pi i\boldsymbol{\eta}^t\boldsymbol{\beta}^{int}). \tag{B10}$$

The integral over $\boldsymbol{\beta}^{int}$ yields the K-dimensional delta function $\delta(\boldsymbol{\eta}-\mathbf{A}^t\boldsymbol{\xi})$, which can then be used to perform the integral over $\boldsymbol{\eta}$. The final result is

$$\mathrm{pr}(\mathbf{g}|\boldsymbol{\alpha},\beta^{ext}) = \int_{\infty}\mathrm{d}^M\boldsymbol{\xi}\exp(-2\pi^2\sigma^2\|\boldsymbol{\xi}\|^2) \times \exp(-2\pi^2\boldsymbol{\xi}^2\mathbf{A}\mathbf{C}\mathbf{A}^t\boldsymbol{\xi})\exp[2\pi i\boldsymbol{\xi}^t(\mathbf{g}-\bar{\mathbf{g}}_0)] = [(2\pi)^M\det(\mathbf{K}_{tot})]^{-1/2} \times \exp\Big\{-\tfrac{1}{2}[\mathbf{g}-\bar{\mathbf{g}}_0]^t\mathbf{K}_{tot}^{-1}[\mathbf{g}-\bar{\mathbf{g}}_0]^t\Big\}, \tag{B11}$$

where $\mathbf{K}_{tot}\equiv\sigma^2\mathbf{I}+\mathbf{A}\mathbf{C}\mathbf{A}^t$.

APPENDIX C: USE OF MAXIMUM-LIKELIHOOD INVARIANCE IN A SHACK–HARTMANN SENSOR

Suppose we have used data from a Shack–Hartmann sensor to obtain ML estimates of tilts. Can we apply the ML invariance principle [Eq. (2.15)] to get ML estimates of the mirror-mode coefficients $\{\alpha_n\}$? The answer is yes if we

can find a matrix $\mathbf{B}$ such that $\boldsymbol{\alpha}=\mathbf{B}\boldsymbol{\tau}$, in which case Eq. (2.15) shows that $\hat{\boldsymbol{\alpha}}_{ML}=\mathbf{B}\hat{\boldsymbol{\tau}}_{ML}$.

To seek such a matrix, we first take the scalar product of Eqs. (4.2) and (4.4) with one of the tilt functions defined in Eq. (4.3); the result is

$$(\chi_k,W)=\sum_{n=1}^{N}\alpha_n(\chi_k,\psi_n)+(\chi_k,\Delta W)=\frac{1}{\|\chi\|^2}\tau_k+(\chi_k,\delta W), \tag{C1}$$

where we have used the orthogonality of the tilt functions, and the division by $\|\chi\|^2$ is needed since the functions were not normalized. (We assume that all lenslets are identical, so that $\|\chi\|^2\equiv(\chi_k,\chi_k)$ is the same for all k.)

As we noted in Subsection 4.A, Eq. (4.4) is an orthogonal decomposition if the region defined by the lenslet is small enough; in that case, $(\chi_k,\delta W)=0$, and we find

$$\tau_k=\sum_{n=1}^{N}M_{kn}\alpha_n+(\chi_k,\Delta W), \tag{C2}$$

where $M_{kn}=(\chi_k,\psi_n)/\|\chi\|^2$.

To proceed, we need to argue that $(\chi_k,\Delta W)=0$, but we cannot do so on the basis of orthogonality. The best we can do is assume that N is large so that the sum in Eq. (4.2) represents the wave exactly and the residual $\Delta W(\mathbf{r})$ is not needed. In that case we have

$$\boldsymbol{\tau}=\mathbf{M}\boldsymbol{\alpha}. \tag{C3}$$

where $\mathbf{M}$ is a $2J\times N$ matrix.

The $N\times N$ matrix $\mathbf{M}^t\mathbf{M}$ will be nonsingular if $2J\geq N$, and the functions $\{\psi_n(\mathbf{r})\}$ are linearly independent, which they always will be in practice. Then we can write

$$\boldsymbol{\alpha}=[\mathbf{M}^t\mathbf{M}]^{-1}\mathbf{M}^t\boldsymbol{\tau}\equiv\mathbf{B}\boldsymbol{\tau}. \tag{C4}$$

To summarize, we can write $\boldsymbol{\alpha}=\mathbf{B}\boldsymbol{\tau}$ only for a high-order AO system (large N) in which all wavefronts of interest are well represented by a linear superposition of mirror influence functions, and then only if the regions defined by the Shack–Hartmann sensor are small and $2J\geq N$.

APPENDIX D: STATISTICS OF CENTROID ESTIMATES IN A SHACK–HARTMANN SENSOR

Traditional data processing in a Shack–Hartmann sensor attempts to estimate the centroids of the irradiance distribution $I(\mathbf{r})$ produced by each lenslet on the detector plane. For simplicity we consider a single lenslet centered on the origin of coordinates, and we delete the index j used to distinguish lenslets.

The centroid location is defined in vector form as

$$\mathbf{r}_c=\frac{\int_{\infty}\mathrm{d}^2r\,\mathbf{r}I(\mathbf{r})}{\int_{\infty}\mathrm{d}^2r\,I(\mathbf{r})}, \tag{D1}$$

where $\mathbf{r}_c\equiv(x_c,y_c)$ is a 2×1 column vector giving the x–y coordinates of the centroid on the detector plane. The traditional centroid estimator is

$$\hat{\mathbf{r}}_c(\mathbf{g})=\frac{1}{g_{tot}}\sum_{m=1}^{M}\mathbf{r}_m g_m, \tag{D2}$$

where $\mathbf{r}_m$ is a 2×1 vector specifying the center location of the mth detector, g_m is the signal from that detector, and g_{tot} is the total signal, given by

$$g_{tot}=\sum_{m=1}^{M}g_m. \tag{D3}$$

A useful way of rewriting Eq. (D2) is

$$\hat{\mathbf{r}}_c(\mathbf{g})=\frac{1}{g_{tot}}\mathbf{R}\mathbf{g}, \tag{D4}$$

where $\mathbf{g}$ is the usual $M\times1$ data vector and $\mathbf{R}$ is a $2\times M$ matrix with elements $R_{km}=x_m$ for $k=1$ and $R_{km}=y_m$ for $k=2$. This form shows that $\hat{\mathbf{r}}_c(\mathbf{g})$ is almost but not quite a linear function of the data $\mathbf{g}$; the linearity is spoiled by the factor $1/g_{tot}$.

From the estimated centroid, an estimate of the 2D tilt vector associated with a given lenslet is traditionally obtained by

$$\hat{\boldsymbol{\tau}}(\mathbf{g})\equiv\hat{\mathbf{r}}_c/z_0, \tag{D5}$$

where z_0 is the distance from the lenslet pupil to the detector plane (usually but not necessarily the focal length). It is hoped (and usually assumed) that $\hat{\boldsymbol{\tau}}(\mathbf{g})$ is an unbiased estimator of the true local tilts $\boldsymbol{\tau}$, that the x and y components of the estimate are uncorrelated Gaussian random variables, and that the estimate is optimal in some sense. The likelihood theory developed in this paper gives us the tools to examine these properties in detail.

A complete treatment of the statistical properties of $\hat{\mathbf{r}}_c$ requires its conditional PDF $\mathrm{pr}(\hat{\mathbf{r}}_c|\boldsymbol{\theta})$, where of course $\boldsymbol{\theta}$ must include all parameters that influence the data. It is convenient to approach this problem by use of the bivariate characteristic function, defined by

$$\Psi_{\hat{\mathbf{r}}_c|\boldsymbol{\theta}}(\boldsymbol{\xi})\equiv\langle\exp[2\pi i\boldsymbol{\xi}^t\hat{\mathbf{r}}_c]\rangle_{\hat{\mathbf{r}}_c|\boldsymbol{\theta}}, \tag{D6}$$

where $\boldsymbol{\xi}$ is a 2×1 vector and the angle brackets indicate expectation with respect to the PDF $\mathrm{pr}(\hat{\mathbf{r}}_c|\boldsymbol{\theta})$. Since $\hat{\mathbf{r}}_c$ is a known function of $\mathbf{g}$, we can equally well perform this expectation with respect to $\mathrm{pr}(\mathbf{g}|\boldsymbol{\theta})$. Using Eq. (D4), we can rewrite Eq. (D6) as

$$\Psi_{\hat{\mathbf{r}}_c|\boldsymbol{\theta}}(\boldsymbol{\xi})=\left\langle\left\langle\exp\left[2\pi i\frac{1}{g_{tot}}\boldsymbol{\xi}^t\mathbf{R}\mathbf{g}\right]\right\rangle_{\mathbf{g}|\boldsymbol{\theta}g_{tot}}\right\rangle_{\mathbf{g}_{tot}|\boldsymbol{\theta}}. \tag{D7}$$

The inner expectation in Eq. (D7) is related to the conditional characteristic function of the data (conditioned on g_{tot} as well as $\boldsymbol{\theta}$), defined by

$$\Psi_{g|\boldsymbol{\theta}g_{tot}}(\boldsymbol{\rho})\equiv\langle\exp[2\pi i\boldsymbol{\rho}^t\mathbf{g}]\rangle_{\mathbf{g}|\boldsymbol{\theta}g_{tot}}, \tag{D8}$$

where $\boldsymbol{\rho}$ is an $M\times1$ vector. Thus,

$$\Psi_{\hat{\mathbf{r}}_c|\boldsymbol{\theta}}(\boldsymbol{\xi})\equiv\left\langle\Psi_{\hat{\mathbf{g}}|\boldsymbol{\theta}g_{tot}}\left(\frac{1}{g_{tot}}\mathbf{R}^t\boldsymbol{\xi}\right)\right\rangle_{\mathbf{g}_{tot}|\boldsymbol{\theta}_t}. \tag{D9}$$

This result shows that the characteristic function (and hence all statistical properties) of the centroid estimates

can be found from the M-dimensional conditional characteristic function of the data by making the substitution indicated in Eq. (D9) and then performing a final one-dimensional average over g_{tot}. If the PDF $\mathrm{pr}(\hat{\mathbf{r}}_c|\boldsymbol{\theta})$ is desired, it can be obtained by performing an inverse 2D Fourier transform.

In two important special cases, the conditional characteristic function of the data can be expressed analytically. If $\mathbf{g}$ follows Poisson statistics without the condition on g_{tot}, then the conditional probability law, for g_{tot} detected photons, is multinomial.[6]

The corresponding conditional characteristic function is[56]

$$\Psi_{\mathbf{g}|\boldsymbol{\theta},g_{tot}}(\boldsymbol{\rho}) = \left[\sum_{m=1}^{M} p_m(\boldsymbol{\theta})\exp(2\pi i\rho_m)\right]^{g_{tot}}, \qquad \text{(D10)}$$

where $p_m(\boldsymbol{\theta})$ is the probablility that a detected photon will be detected in the mth detector element: $p_m(\boldsymbol{\theta}) = \bar{g}_m(\boldsymbol{\theta})/\bar{g}_{tot}$.

If $\mathbf{g}$ follows a multivariate normal law without the condition on g_{tot}, then the conditional PDF is also normal, and the requisite conditional mean and covariance matrix can be found from Eqs. (B4) and (B5), respectively. The final average over g_{tot} spoils the normal character of the centroid statistics, however, even with normally distributed data.

No analytic form for the final characteristic function of the centroid estimates has been found for either the Poisson or the normal case, but the average is easily performed numerically since it is one-dimensional.

ACKNOWLEDGMENTS

We thank Nicholas Devaney, Thomas Farrell, Lars Furenlid, and Jacob Hesterman for many stimulating discussions. We also thank Richard Lane for prompting us to consider the effects of correlations induced by marginalization. This research was supported by Science Foundation Ireland (SFI) under grant 01/PI.2/B039C and by an SFI Walton Fellowship (03/W3/M420) for H. H. Barrett. Related methods in nuclear medicine were supported in part by the National Institutes of Health under grant P41 EB002035.

Corresponding author H.H. Barrett can be reached by e-mail at hhb@email.arizona.edu; C. Dainty, c.dainty @nuigalway.ie; D. Lara, d.lara@nuigalway.ie

REFERENCES

1. R. K. Tyson, *Principles of Adaptive Optics* (Academic Press, 1998).
2. G. Rousset, "Wavefront sensing," in *Adaptive Optics in Astronomy*, F. Roddier, ed. (Cambridge U. Press, 1999).
3. F. Roddier, "Curvature sensing and compensation: a new concept in adaptive optics," Appl. Opt. **27**, 1223–1225 (1988).
4. G. A. Tyler and D. L. Fried, "Image-position error associated with a quadrant detector," J. Opt. Soc. Am. **72**, 804 (1982).
5. E. P. Wallner, "Optimal wave-front correction using slope measurements," J. Opt. Soc. Am. **73**, 1771–1776 (1983)
6. H. H. Barrett and K. J. Myers, *Foundations of Image Science* (Wiley, 2004).
7. H. H. Barrett, "Objective assessment of image quality: effects of quantum noise and object variability," J. Opt. Soc. Am. A **7**, 1266–1278 (1990).
8. H. H. Barrett, J. L. Denny, R. F. Wagner, and K. J. Myers, "Objective assessment of image quality: II. Fisher information, Fourier crosstalk, and figures of merit for task performance," J. Opt. Soc. Am. A **12**, 834–852 (1995).
9. H. H. Barrett, C. K. Abbey, and E. Clarkson, "Objective assessment of image quality: III. ROC metrics, ideal observers and likelihood-generating functions," J. Opt. Soc. Am. A **15**, 1520–1535 (1998).
10. B. E. A. Saleh, "Estimation of the location of an optical object with photodetectors limited by quantum noise," Appl. Opt. **13**, 1824–1827 (1974).
11. B. E. A. Saleh, "Estimations based on instants of occurrence of photon counts of low level light," Proc. IEEE **62**, 530–531 (1974).
12. B. E. A. Saleh, "Joint probability of occurrence of photon events and estimation of optical parameters," J. Phys. A **7**, 1360–1368 (1974).
13. M. Elbaum and M. Greenebaum, "Annular apertures for angular tracking," Appl. Opt. **16**, 2438–2440 (1977).
14. K. A. Winick, "Cramér–Rao lower bounds on the performance of charge-coupled-device optical position estimator," J. Opt. Soc. Am. A **3**, 1809–1815 (1986).
15. R. Irwan and R. G. Lane, "Analysis of optimal centroid estimation applied to Shack–Hartmann sensing," Appl. Opt. **38**, 6737–6743 (1999).
16. M. A. van Dam, "Wave-front sensing for adaptive optics in astronomy," Ph. D. thesis (University of Canterbury, 2002).
17. M. A. van Dam and R. G. Lane, "Wave-front slope estimation," J. Opt. Soc. Am. A **17**, 1319–1324 (2000).
18. B. M. Welsh, B. L. Ellerbroek, M. C. Roggemann, and T. L. Pennington, "Fundamental performance comparison of a Hartmann and a shearing interferometer wavefront sensor," Appl. Opt. **34**, 4186–4195 (1995).
19. M. G. Löfdahl, A. L. Duncan, and G. B. Scharmer, "Fast-phase diversity wave-front sensing for mirror control," in Proc. SPIE **3353**, 952–963 (1988).
20. S. A. Sallberg, B. M. Welsh, and M. C. Roggemann, "Maximum *a posteriori* estimation of wave-front slopes using a Shack–Hartmann wave-front sensor," J. Opt. Soc. Am. A **14**, 1347–1354 (1997).
21. R. C. Cannon, "Global wave-front reconstruction using Shack–Hartmann sensors," J. Opt. Soc. Am. A **12**, 2031–2039 (1995).
22. A. Blanc, L. M. Mugnier, and J. Idier, "Marginal estimation of aberrations and image restoration by use of phase diversity," J. Opt. Soc. Am. A **20**, 1035–1045 (2003).
23. R. A. Gonsalves, "Phase retrieval and diversity in adaptive optics," Opt. Eng. **21**, 829–832 (1982).
24. R. G. Paxman, T. J. Schulz, and J. R. Fienup, "Joint estimation of object and aberrations using phase diversity," J. Opt. Soc. Am. A **7**, 1072–1085 (1992).
25. J. J. Dolne, R. J. Tansey, K. A. Black, J. H. Deville, P. R. Cunningham, K. C. Widen, and P. S. Idell, "Practical issues in wave-front sensing by use of phase diversity," Appl. Opt. **42**, 5284–5289 (2003).
26. R. M. Gray and A. Macovski, "Maximum *a posteriori* estimation of position in scintillation cameras," IEEE Trans. Nucl. Sci. **NS-23**, 849–852 (1976).
27. J. N. Aarsvold, H. H. Barrett, J. Chen, A. L. Landesman, T. D. Milster, D. D. Patton, T. J. Roney, R. K. Rowe, R. H. Seacat III, and L. M. Strimbu, "Modular scintillation cameras: a progress report," in Proc. SPIE **914**, 319–325 (1988).
28. T. D. Milster, J. N. Aarsvold, H. H. Barrett, A. L. Landesman, L. S. Mar, D. D. Patton, T. J. Roney, R. K. Rowe, and R. H. Seacat III, "A full-field modular gamma camera," J. Nucl. Med. **31**, 632–639 (1990).
29. D. Gagnon, "Maximum likelihood positioning in the scintillation camera using depth of interaction," IEEE Trans. Med. Imaging **MI-12**, 101–107 (1993).
30. N. H. Clinthorne, W. L. Rogers, L. Shao, and K. F. Kcral, "A hybrid maximum likelihood position computer for

scintillation cameras," IEEE Trans. Nucl. Sci. **34**, 97–101 (1987).
31. L. R. Furenlid, J. Y. Hesterman, and H. H. Barrett, "Real time data acquisition and maximum-likelihood estimation for gamma cameras," in *Proceedings of the 14th IEEE–NPSS Real-Time Conference* (IEEE, 2005), pp. 498–501.
32. J. L. Melsa and D. L. Cohn, *Decision and Estimation Theory* (McGraw-Hill, 1978).
33. H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, Vol. I (Wiley, 1968).
34. L. L. Scharf, *Statistical Signal Processing: Detection, Estimation, and Time-Series Analysis* (Addison-Wesley, 1991).
35. H. H. Barrett, K. J. Myers, N. Devaney, and J. C. Dainty, "Objective assessment of image quality IV. Application to adaptive optics," J. Opt. Soc. Am. A 3080–3105 (2006).
36. P. Stoica and T. L. Marzetta, "Parameter estimation problems with singular information matrices," IEEE Trans. Signal Process. **49**, 87–89 (2001).
37. J. O. Berger, B. Liseo, and R. L. Wolpert, "Integrated likelihood methods for eliminating nuisance parameters," Stat. Sci. **14**, 1–28 (1999).
38. H. Cramér, *Mathematical Methods of Statistics* (Princeton U. Press, 1946).
39. H. H. Barrett, L. Parra, and T. A. White, "List-mode likelihood," J. Opt. Soc. Am. A **14**, 2914–2923 (1997).
40. L. Parra and H. H. Barrett, "List-mode likelihood: EM algorithm and noise estimation demonstrated on 2D-PET," IEEE Trans. Med. Imaging **MI-17**, 228–235 (1998).
41. D. L. Snyder, C. W. Helstrom, A. D. Lanterman, and M. Faisal, "Compensation for readout noise in CCD images," J. Opt. Soc. Am. A **12**, 272–283 (1995).
42. B. E. A. Saleh and M. C. Teich, "Multiplied Poisson noise in pulse, particle, and photon detection," Proc. IEEE **70**, 229–245 (1992).
43. B. W. Miller, H. B. Barber, H. H. Barrett, I. Shestakova, B. Singh, and V. V. Nagarkar, "Single-photon spatial resolution enhancement of columnar CsI(Tl) using centroid estimation and event discrimination," Proc. SPIE **6142**, 61421T (2006).
44. R. E. Burgess, "Homophase and heterophase fluctuations in semiconducting crystals," Discuss. Faraday Soc. **21**, 51–158 (1959).
45. R. K. Swank, "Absorption and noise in X-ray phosphors," J. Appl. Phys. **44**, 4199–4203 (1973).
46. M. Rabbani, R. Shaw, and R. van Metter, "Detective quantum efficiency of imaging systems with amplifying and scattering mechanisms," J. Opt. Soc. Am. A **4**, 895–901 (1987).
47. H. H. Barrett, R. F. Wagner, and K. J. Myers, "Correlated point processes in radiological imaging," in Proc. SPIE **3032**, 110–124 (1997).
48. L. Chen and H. H. Barrett, "Non-Gaussian noise in X-ray and gamma-ray detectors," in Proc. SPIE **5745**, 366–376 (2005).
49. R. G. Paxman, H. H. Barrett, W. E. Smith, and T. D. Milster, "Image reconstruction from coded data: II. Code design," J. Opt. Soc. Am. A **2**, 501–509 (1985).
50. H. White, "Maximum likelihood estimation of misspecified models," Econometrica, **50**, 1–126 (1982).
51. R. J. Noll, "Zernike polynomials and atmospheric turbulence," J. Opt. Soc. Am. **66**, 207–210 (1976).
52. H. H. Barrett, "Detectors for small-animal SPECT: II. Statistical limitations and estimation methods," in *Small-Animal SPECT Imaging*, M. Kupinski and H. Barrett eds. (Springer, 2005), Chap. 3.
53. J. Y. Hesterman (University of Arizona, jyh@email.arizona.edu, personal communication, 2005).
54. S. T. Smith, "Covariance, subspace, and intrinsic Cramer-Rao bounds," IEEE Trans. Signal Process. **53**, 1610–1630 (2005).
55. K. V. Mardia, J. T. Kent, and J. M. Bibby, *Multivariate Analysis* (Academic, 1979).
56. N. L. Johnson and S. Kotz, *Discrete Distributions* (Wiley, 1969).