Mean and Variance of Implicitly Defined Biased Estimators (Such as Penalized Maximum Likelihood): Applications to Tomography

Jeffrey A. Fessler, Member, IEEE

Abstract—Many estimators in signal processing problems are defined implicitly as the maximum of some objective function. Examples of implicitly defined estimators include maximum likelihood, penalized likelihood, maximum a posteriori, and nonlinear least squares estimation. For such estimators, exact analytical expressions for the mean and variance are usually unavailable. Therefore, investigators usually resort to numerical simulations to examine properties of the mean and variance of such estimators. This paper describes approximate expressions for the mean and variance of implicitly defined estimators of unconstrained continuous parameters. We derive the approximations using the implicit function theorem, the Taylor expansion, and the chain rule. The expressions are defined solely in terms of the partial derivatives of whatever objective function one uses for estimation. As illustrations, we demonstrate that the approximations work well in two tomographic imaging applications with Poisson statistics. We also describe a "plug-in" approximation that provides a remarkably accurate estimate of variability even from a single noisy Poisson sinogram measurement. The approximations should be useful in a wide range of estimation problems.

I. INTRODUCTION

Let \( \theta = [\theta_1, \ldots, \theta_p] \in \mathbb{R}^p \) be an unknown real parameter vector that is to be estimated from a measurement vector \( Y = [Y_1, \ldots, Y_N] \in \mathbb{R}^N \), where \( \cdot \) denotes vector or matrix transpose. In many areas of signal and image processing, one specifies an estimator \( \hat{\theta} \) to be the maximum of some objective function:

\[
\hat{\theta} = \arg \max_\theta \Phi(\theta, Y). \tag{1}
\]

Examples of such methods include maximum-likelihood estimation, maximum a posteriori or penalized-likelihood methods, and linear or nonlinear least squares methods. Except in very simple cases such as linear least squares estimation, there is usually no analytical form that explicitly expresses \( \hat{\theta} \) in terms of \( Y \). In other words, the objective function (1) only \textit{implicitly} defines \( \hat{\theta} \) as a function of \( Y \). Statisticians refer to (1) as an \textit{M-estimate} [1].

The absence of an explicit analytical expression of the form \( \hat{\theta} = h(Y) \) makes it difficult to study the mean and variance of the estimator \( \hat{\theta} \), except through numerical simulations. Often the estimators of interest depend on one or more “tuning parameters,” such as the regularization parameter in penalized-likelihood methods, and one would like to be able to easily study the estimator characteristics over a range of values for those parameters. In such cases, numerical simulations can be prohibitively expensive for complicated estimators (particularly when \( p \) is large). Similar considerations apply if one wishes to compare estimator performance against the uniform Cramer–Rao bound for biased estimators [2], [3] to examine the bias–variance tradeoff of the estimator. Therefore, it would be useful to have approximate expressions for the mean and variance of implicitly defined estimators, particularly if those approximations require less computation than multiple numerical simulations [4].

For unbiased maximum-likelihood estimation, the Cramer–Rao bound can serve as an approximation to the estimator variance. Our focus is on regularized methods for which bias is unavoidable, so the unbiased Cramer–Rao bound is inapplicable. Approximate covariances for penalized-likelihood estimates have been computed for specific iterative algorithms [5], but most analyses of penalized-likelihood methods have focussed on the asymptotic properties of mean squared error, e.g., [6], [7]. For practical signal-to-noise ratios, bias and variance may have unequal importance, in contrast to their equal weighting in the mean-squared-error performance measure.

In this paper we apply the implicit function theorem, the Taylor expansion, and the chain rule to (1) to derive approximate expressions for the mean and variance of implicitly defined estimators \( \hat{\theta} \). Evaluating these expressions numerically typically requires a similar amount of computation as one or two realizations in a numerical simulation. Therefore, these expressions allow one to quickly determine “interesting” values for the tuning parameters, etc., for further investigation using numerical simulations. In addition, one can use the variance approximation to determine how many realizations are needed to achieve a desired accuracy in subsequent numerical simulations.

Our expressions are similar to the asymptotic moments given by Serfling [1] for scalar \( M \)-estimates. Our focus here is
on presenting a simple derivation of useful approximations for multiparameter imaging problems, rather than on asymptotics. The Appendix compares in more detail the two approaches.

Because of the partial derivatives used in the derivation, our approximations are restricted to problems where is a continuous parameter. Thus the approach is inapplicable to discrete classification problems such as image segmentation. (Mean and variance are poor performance measures for segmentation problems anyway; analyses of classification errors are more appropriate [8].) Furthermore, strictly speaking we must also exclude problems where inequality constraints are imposed on , since when the maximization in (1) is subject to inequality constraints, one must replace (2) below with appropriate Karush-Kuhn-Tucker conditions. Our focus is on imaging problems, where often the only inequality constraint is nonnegativity of . Although this constraint is often important in unpenalized estimation methods, our primary interest is in objective functions () that include a regularization term. In our experience, the nonnegativity constraints are active relatively infrequently with regularized estimates, so the variances of the unconstrained and constrained estimators are approximately equal for most pixels (cf [9]). We demonstrate this property empirically in Section IV, where the mean and variance approximation for the unconstrained estimator agree closely with the empirical performance of an estimator implemented with nonnegativity constraints.

Our derivation assumes the estimate is computed by “completely” maximizing an objective function, i.e., the approximations are not applicable to unregularized objective functions for which one uses a “stopping rule” to terminate the iterations long before the maximum is reached. In particular, our results are inapplicable to unregularized methods such as iterative filtered backprojection [10], the ordered subsets expectation maximization algorithm [11], or weighted least squares conjugate gradient [12]. Except in simple linear cases [13], it is generally difficult to analyze the performance of methods based on stopping rules, although Barrett et al. [14], [15] have analyzed the per-iteration behavior of the maximum-likelihood expectation maximization algorithm for emission tomography. The approximations we derive are somewhat easier to use since they are independent of the number of iterations (provided sufficient iterations are used to maximize the objective function).

Section II develops the mean and variance approximations. We expect these approximations to be useful in many types of signal processing problems. However, the particular tradeoffs between the cost of the computing the approximations and the cost of performing numerical simulations will likely differ between applications. Therefore, we devote most of the paper to concrete illustrations of the utility and accuracy of the approximations on two tomographic imaging applications. Section III describes the (linear) regularized least squares estimator. Section IV illustrates that the approximations are accurate even for a highly nonlinear penalized-likelihood estimator in a transmission tomographic imaging application. Section V illustrates how one can use the variance approximation to obtain remarkably accurate estimates of variance even from a single noisy measurement (e.g., real data) using a simple plug-in approach. Section VI describes an emission tomographic imaging application, where we show that a penalized least squares estimator has a systematic bias at low count rates.

II. APPROXIMATIONS

We assume () has a unique global maximum that is well defined. We also restrict our attention to suitably regular objective functions for which one can find the required maximum in (1) by zeroing the partial derivatives of ()

\[ 0 = \frac{\partial}{\partial \theta_j} \Phi(\theta, Y) \bigg|_{\theta = \hat{\theta}}, \quad j = 1, \ldots, p. \] (2)

It is this assumption that restricts our approximations to continuous parameters and that precludes inequality constraints and stopping rules.

For suitably regular , the assumption of uniqueness and the implicit function theorem [16, p. 266] ensure that the relationship (2) implicitly defines a function \( \hat{\theta} = h(Y) = (h_1(Y) \cdots h_p(Y))^\top \) that maps the measurement into an estimate . From (2) the function \( h(Y) \) must satisfy

\[ 0 = \frac{\partial}{\partial \theta_j} \Phi(h(Y), Y), \quad j = 1, \ldots, p \] (3)

With perhaps a slight abuse of notation, we will rewrite (3) as

\[ 0 = \frac{\partial}{\partial \theta_j} \Phi(h(Y), Y), \quad j = 1, \ldots, p \] (4)

where we will always use \( \frac{\partial}{\partial \theta_j} \) to denote partial derivatives with respect to the first argument of the function \( \Phi(\theta, Y) \), and \( \frac{\partial}{\partial Y_k} \) to denote partial derivatives with respect to the second argument, regardless of what values are used to evaluate the resulting derivatives.

The implicitly defined function \( h(Y) \) can rarely be found analytically, and one usually implements an iterative method for maximizing \( \Phi(\cdot, Y) \) to find . Even if one did have an analytical expression for \( h(Y) \), it would still be difficult to compute its mean or variance exactly since the estimator \( h(Y) \) is usually nonlinear. Although exact analytical expressions for the mean and variance of \( h(Y) \) are unavailable, if we knew \( h(Y) \) we could approximate its mean and variance using standard methods based on the second-order Taylor expansion of \( h(Y) \). If \( Y_n \) denotes the mean of \( Y_n \), then

\[ h(Y) \approx h(Y) + \sum_n \frac{\partial}{\partial Y_n} h(Y)(Y_n - \bar{Y}_n) \]

\[ + \frac{1}{2} \sum_n \sum_m \frac{\partial^2}{\partial Y_n \partial Y_m} h(Y)(Y_n - \bar{Y}_n)(Y_m - \bar{Y}_m). \] (5)

We use this expansion in the following to derive approximations for the covariance and mean of \( \hat{\theta} = h(Y) \).
A. Covariance

For the covariance approximation we use the first-order Taylor expansion in matrix form

\[ h(Y) \approx \hat{h}(Y) + \nabla h(Y)(Y - \bar{Y}) \]  

(6)

where \( \nabla = \left[ \frac{\partial}{\partial Y_1}, \ldots, \frac{\partial}{\partial Y_N} \right] \) denotes the (row) gradient operator. Taking the covariance of both sides yields the following well-known approximation [17, p. 426]:

\[ \text{Cov} \{ \hat{\theta} \} = \text{Cov} \{ h(Y) \} \approx \nabla h(Y) \text{Cov} \{ Y \} \nabla h(Y)^\top. \]  

(7)

If we knew \( h(Y) \) then we could directly apply (7) to approximate the covariance of \( \hat{\theta} = h(Y) \). But since \( h(Y) \) is unknown, (7) is not immediately useful. However, the dependence on \( h(Y) \) in (7) is only through its partial derivatives at the point \( \bar{Y} \). From the calculus of vector functions [18, p. 302], one can determine the partial derivatives of an implicitly defined function by applying the chain rule. Differentiating (4) with respect to \( Y_n \) by applying the chain rule \( ^2 \) yields

\[ 0 = \sum_k \frac{\partial^2}{\partial \theta_j \partial \theta_k} \Phi(h(Y), Y) \frac{\partial h_k(Y)}{\partial Y_n} + \frac{\partial^2}{\partial \theta_j \partial Y_n} \Phi(h(Y), Y) \]  

(8)

\( j = 1, \ldots, p, n = 1, \ldots, N \). This equality gives \( N \) sets of \( p \) equations in \( p \) unknowns, and it holds for any \( \bar{Y} \). However, since (7) only depends on \( \nabla h(Y) \), we only need the special case of (8) where \( \bar{Y} = \bar{Y} \). Writing that case in matrix form

\[ 0 = \nabla^{20} \Phi(h(\bar{Y}), \bar{Y}) \nabla h(\bar{Y}) + \nabla^{11} \Phi(h(\bar{Y}), \bar{Y}) \]  

(9)

where the \((j,k)\)th element of the \( p \times p \) operator \( \nabla^{20} \) is \( \frac{\partial^2 \Phi}{\partial \theta_j \partial \theta_k} \), and the \((j,n)\)th element of the \( p \times N \) operator \( \nabla^{11} \) is \( \frac{\partial^2 \Phi}{\partial \theta_j \partial Y_n} \). To proceed, we now assume that the symmetric matrix \( -\nabla^{20} \Phi(h(\bar{Y}), \bar{Y}) \) is also positive definite, \(^3 \) so we can solve for \( \nabla h(Y) \) by rearranging

\[ \nabla h(Y) = \left[ -\nabla^{20} \Phi(h(\bar{Y}), \bar{Y}) \right]^{-1} \nabla^{11} \Phi(h(\bar{Y}), \bar{Y}). \]  

(10)

If we define \( \hat{\theta} = h(Y) \), then combining (10) with (7) yields the following approximation for the covariance of \( \hat{\theta} \):

\[ \text{Cov} \{ \hat{\theta} \} \approx \left[ -\nabla^{20} \Phi(h(\bar{Y}), \bar{Y}) \right]^{-1} \nabla^{11} \Phi(h(\bar{Y}), \bar{Y}) \text{Cov} \{ Y \} \left[ -\nabla^{20} \Phi(h(\bar{Y}), \bar{Y}) \right]^{-1}. \]  

(11)

When \( p \) is large, storing the full covariance matrix is inconvenient, and often one is interested primarily in the variance of certain parameters in a region of interest. Let \( e^j \) be the \( j \)th unit vector of length \( p \), and define \( u^j = [-\nabla^{20} \Phi(h(\bar{Y}), \bar{Y})]^{-1} e^j \). Note that one does not need to perform a \( p \times p \) matrix inversion to compute \( u^j \); one simply solves the equation \([-\nabla^{20} \Phi(h(\bar{Y}), \bar{Y})]u^j = e^j \), which can be done directly when \( p \) is small, or via fast iterative methods such as Gauss–Siedel when \( p \) is large [19]. From (11) it follows that

\[ \text{Cov} \{ \hat{\theta}_j, \hat{\theta}_k \} = (e^j)^\top \text{Cov} \{ \hat{\theta} \} e^k \approx (u^j)^\top \nabla^{11} \Phi(h(\bar{Y}), \bar{Y}) \text{Cov} \{ Y \} \nabla^{11} \Phi(h(\bar{Y}), \bar{Y})^\top u^k \]  

(12)

for \( j, k = 1, \ldots, p \). One can compute any portion of the covariance matrix of \( \hat{\theta} \) by using (12) repeatedly for appropriate \( j \) and \( k \). In general, computing \( \text{Var} \{ \hat{\theta}_j \} \) using this formula requires \( O(p^2 + np + n^2) \) operations. In many problems, such as the tomographic examples in Sections IV and VI, the covariance of \( Y \) is diagonal and the partial derivatives have a sparse structure, so the actual computation is much less.

To summarize, (11) and (12) are the main results of this subsection: approximate expressions for the estimator (co)variance that depend only \(^4 \) on the partial derivatives of the objective function \( \Phi(\theta, Y) \), and do not require an expression for the implicit function \( h(Y) \).

B. Mean

To approximate the mean of \( \hat{\theta} = h(Y) \) one has two choices. The simplest approach is to take the expectation of the 0th-order Taylor expansion, yielding the approximation

\[ E\{ \hat{\theta} \} = E\{ h(Y) \} \approx h(\bar{Y}) = \hat{\theta}. \]  

(13)

This approximation is simply the value produced by applying the estimator (1) to noise-free data. This approach requires minimal computation, and works surprisingly well for penalized-likelihood objectives. It has been used extensively by investigators in emission tomography [14], [15], [20]. Apparently, the principal source of bias in penalized-likelihood estimators is the regularizing penalty that one includes in the equation \( \Phi(\theta, Y) \), and we next derive a mean approximation based on the second-order Taylor expansion, which is more accurate, but has the disadvantage of greater computation.

Taking the expectation of both sides of the second-order Taylor expansion (5) yields the following well-known approx-
imation for the mean of $h(Y)$:

$$ E\{\hat{\theta}\} \approx h(Y) + \frac{1}{2} \sum_n \sum_m \frac{\partial^2}{\partial Y_n \partial Y_m} h(Y) \text{Cov}\{Y_n, Y_m\}, $$

(14)

where

$$ \text{Cov}\{Y_n, Y_m\} = E\{(Y_n - \bar{Y}_n)(Y_m - \bar{Y}_m)\} $$

is the $(n,m)$th element of the covariance matrix of $Y$. The approximation (14) requires the second partial derivatives of $h(Y)$. To obtain those partial derivatives, we use the chain rule to differentiate (8) again with respect to $U$, obtaining

$$ \frac{\partial^2}{\partial \theta_j \partial \theta_k} \Phi(h(Y), Y) = \frac{\partial}{\partial Y_k} h(Y), $$

(15)

for $j = 1, \ldots, p, n = 1, \ldots, N, m = 1, \ldots, N$. One can substitute in $\theta = h(Y)$ and $Y = Y$ in the above expression to obtain $N^2$ sets of $p$ equations in the $p$ unknowns $\{\frac{\partial}{\partial Y_k} h(Y)\}_{k=1}^p$. Solving each of those systems of equations and then substituting back into (14) yields an approximation to $E\{\hat{\theta}\}$ that is independent of the unknown implicit function $h(Y)$. If $p$ and $n$ are large in a given problem, then one must weigh the relative computational expense of solving the above equations versus performing numerical simulations. The tradeoff will depend on the structure of the objective function $\Phi$. Note that (15) depends on the first partials $\frac{\partial}{\partial Y_k} h(Y)$, so one must first apply (10) to compute those partials.

Unlike expression (8), which we were able to write in the matrix form (9), there does not appear to be a simple form for rewriting (15), except by introducing tensor products (which really do not offer much simplification). However, the equations in (15) do simplify for some special cases for $\Phi$, described next.

C. Independent Measurements

If the measurements $Y_1, \ldots, Y_N$ are statistically independent, then (14) simplifies to

$$ E\{\hat{\theta}\} = E\{h(Y)\} \approx h(Y) + \frac{1}{2} \sum_n \frac{\partial^2}{\partial Y_n^2} h(Y) \text{Var}\{Y_n\}, $$

(16)

This expression depends only on the diagonal elements of the covariance of $Y$ and on the diagonal of the matrix of second partial derivatives of $h(Y)$. Therefore, one needs only the cases where $m = n$ in (15), i.e., one needs to solve $N$ sets of $p$ equations in $p$ unknowns of the form

$$ 0 = \sum_k \left[ \sum_\ell \frac{\partial^3}{\partial \theta_\ell \partial \theta_k \partial \theta_l} \Phi(h(Y), Y) \frac{\partial}{\partial Y_k} h(Y) \right] + \frac{\partial^2}{\partial \theta_k \partial \theta_l} \Phi(h(Y), Y) \frac{\partial^2}{\partial Y_k \partial Y_l} h(Y) $$

$$ + \frac{\partial^2}{\partial \theta_k \partial \theta_l} \Phi(h(Y), Y) \frac{\partial^2}{\partial Y_k \partial Y_l} h(Y) $$

$$ + \frac{\partial^3}{\partial \theta_k \partial \theta_l \partial \theta_m} \Phi(h(Y), Y) \frac{\partial}{\partial Y_k} h(Y), \quad j = 1, \ldots, p, n = 1, \ldots, N. $$

(17)

D. Scalar Parameter

If $p = 1$, i.e., $\theta$ is a scalar, then (15) simplifies to

$$ 0 = \left[ \frac{\partial^3}{\partial \theta_\ell \partial \theta_k \partial \theta_l} \Phi(h(Y), Y) \frac{\partial}{\partial Y_k} h(Y) \right] + \frac{\partial^2}{\partial \theta_k \partial \theta_l} \Phi(h(Y), Y) \frac{\partial^2}{\partial Y_k \partial Y_l} h(Y) $$

$$ + \frac{\partial^2}{\partial \theta_k \partial \theta_l} \Phi(h(Y), Y) \frac{\partial^2}{\partial Y_k \partial Y_l} h(Y) $$

$$ + \frac{\partial^3}{\partial \theta_k \partial \theta_l \partial \theta_m} \Phi(h(Y), Y), \quad n = 1, \ldots, N, m = 1, \ldots, N. $$

(17)

By rearranging we can solve explicitly for the second partials of $h(Y)$

$$ \frac{\partial^2}{\partial Y_n \partial Y_m} h(Y) = \left[ -\frac{\partial^2}{\partial \theta^2} \Phi(\bar{\theta}, \bar{Y}) \right]^{-1} \left[ \frac{\partial^3}{\partial \theta_\ell \partial \theta_k \partial \theta_l} \Phi(\bar{\theta}, \bar{Y}) \frac{\partial}{\partial Y_k} h(\bar{Y}) \right] + \frac{\partial^2}{\partial \theta_k \partial \theta_l} \Phi(\bar{\theta}, \bar{Y}) \frac{\partial^2}{\partial Y_k \partial Y_l} h(\bar{Y}) $$

$$ + \frac{\partial^2}{\partial \theta_k \partial \theta_l} \Phi(\bar{\theta}, \bar{Y}) \frac{\partial^2}{\partial Y_k \partial Y_l} h(\bar{Y}) $$

$$ + \frac{\partial^3}{\partial \theta_k \partial \theta_l \partial \theta_m} \Phi(\bar{\theta}, \bar{Y}) \frac{\partial}{\partial Y_k} h(\bar{Y}) + \frac{\partial^2}{\partial \theta_k \partial \theta_l} \Phi(\bar{\theta}, \bar{Y}) \frac{\partial^2}{\partial Y_k \partial Y_l} h(\bar{Y}) + \frac{\partial^3}{\partial \theta_k \partial \theta_l \partial \theta_m} \Phi(\bar{\theta}, \bar{Y}) \right]. $$

Substituting this expression into (14) yields the approximate mean for a scalar parameter estimator.

III. EXAMPLE: REGULARIZED LEAST SQUARES

The approximations for mean and covariance derived above are exact in the special case where the estimator is linear, since in that case the first-order Taylor expansion (6) is exact. In this section we verify this property by computing (11) and (15) for a regularized least squares problem. The expressions are useful for making comparisons with the corresponding approximation for nonlinear estimators derived in the subsequent sections.

Suppose the measurements obey the standard linear model with additive noise

$$ Y = A\theta + \text{noise} $$

where $A$ is a known $N \times p$ matrix. For such problems, the following regularized weighted least squares objective
function is often used for estimation:
\[
\Phi(\theta, Y) = \frac{1}{2} (Y - \mathbf{A}\theta)'\mathbf{R}(\theta)(Y - \mathbf{A}\theta) - \beta R(\theta)
\]
where \( \mathbf{R}(\theta) \) is a nonnegative definite weighting matrix and \( R(\theta) \) is a roughness penalty of the form
\[
R(\theta) = \sum_j \sum_k w_{jk} \phi'(\theta_j - \theta_k)
\]
where \( w_{jk} = 1 \) for horizontal and vertical neighbors, \( w_{jk} = 1/\sqrt{2} \) for diagonal neighbors, and 0 otherwise. Note that
\[
\nabla^2 R(\theta) = 0
\]
and define
\[
\mathbf{R}(\theta) = \nabla^2 R(\theta)
\]
to be the matrix of second partials of \( R(\theta) \). The \((j,k)\)th element of \( R(\theta) \) is
\[
\begin{cases}
\sum_k w_{jk} \phi'(\theta_j - \theta_k), & j = k \\
-w_{jk} \phi'(\theta_j - \theta_k), & j \neq k
\end{cases}
\]
where \( \phi'' \) denotes the second derivative of \( \phi \).

Consider the quadratic case where \( \phi(x) = x^2/2 \), so \( R(\theta) = \frac{1}{2} \theta' \mathbf{R} \theta \). Assume \( \mathbf{R} \) is a symmetric nonnegative definite regularization matrix whose null space is disjoint from the null space of \( \mathbf{A} \mathbf{I} \mathbf{A} \). In this case, one can derive an explicit expression for the estimator
\[
\hat{\theta} = \hat{h}(Y) = (\mathbf{A}' \mathbf{R} \mathbf{I} \mathbf{A} + \beta \mathbf{R})^{-1} \mathbf{A}' \mathbf{R} \mathbf{I} Y
\]
from which one can derive exact expressions for the mean and covariance. However, for didactic purposes, we instead derive the mean and covariance using the “approximations” (11) and (15).

The partial derivatives of \( \Phi(\theta, Y) \) are
\[
\begin{align*}
-\nabla^2 \Phi &= \mathbf{A}' \mathbf{R} \mathbf{I} \mathbf{A} + \beta \mathbf{R} \\
\nabla^3 \Phi &= -\mathbf{A}' \mathbf{R} \\
\nabla^4 \Phi &= \nabla^2 \mathbf{A}' \mathbf{R} \\
\nabla^5 \Phi &= \nabla^3 \Phi = \nabla^5 \Phi = 0
\end{align*}
\]
so substituting into (15), one finds that \( \nabla^2 \hat{h}(Y) = 0 \). Thus from (14)
\[
E\{\hat{\theta} \} = h(Y) = (\mathbf{A}' \mathbf{R} \mathbf{I} \mathbf{A} + \beta \mathbf{R})^{-1} \mathbf{A}' \mathbf{R} \mathbf{I} Y
\]
which of course is exactly what one would get from (20). Substituting (21) into (11) yields the estimator covariance
\[
\text{Cov}\{\hat{\theta} \} = [\mathbf{A}' \mathbf{R} \mathbf{I} \mathbf{A} + \beta \mathbf{R}]^{-1} \mathbf{R} \mathbf{I} \text{Cov}\{Y \} \mathbf{I} \mathbf{A} [\mathbf{A}' \mathbf{R} \mathbf{I} \mathbf{A} + \beta \mathbf{R}]^{-1}
\]
which again agrees with (20). If the measurement covariance is known, then usually one chooses \( \mathbf{I}^{-1} = \text{Cov}\{Y \} \), in which case
\[
\text{Cov}\{\hat{\theta} \} = (\mathbf{F} + \beta \mathbf{R})^{-1} \mathbf{F} (\mathbf{F} + \beta \mathbf{R})^{-1}
\]
functions of the form
\[
\Phi(\theta, Y) = \frac{1}{T} L(\theta, Y) - \beta R(\theta)
\]  
(25)
where the roughness penalty $R$ was defined in (18).

Due to the nonlinearity of (23) and the nonquadratic likelihood function (24) for Poisson statistics, the estimate $\hat{\theta}$ formed by maximizing (25) is presumably a very nonlinear function of $Y$. Furthermore, since attenuation coefficients are nonnegative, one usually enforces the inequality constraint $\hat{\theta} \geq 0$. Therefore, this problem provides a stringent test of the accuracy of the mean and variance approximations.

**A. Covariance Approximation**

Since the number of measurements (or rays) $N$ and the number of parameters (pixels) $p$ are both large, we would like to approximate the variance of certain pixels of interest using (12), which requires the following partial derivatives:

\[
\frac{\partial}{\partial \theta_j} L(\theta, Y) = T \sum_n a_{nj} \left( 1 - \frac{Y_n}{Y_n(\theta)} \right) (p_n(\theta) - r_n)
\]

\[
\frac{\partial^2}{\partial \theta_j \partial \theta_k} L(\theta, Y) = T \sum_n a_{nj} a_{nk} q_n(\theta)
\]

where $q_n(\theta) = \left( 1 - \frac{r_n Y_n / T}{p_n(\theta)} \right) b_n e^{-\sum_j a_{nj} \theta_j}$. 

Combining the above expressions in matrix form with the expressions for the partials of $R$ given in Section III

\[
\nabla^2 \Phi(\theta, Y) = A' \text{diag}(p_n(\theta)) A + \beta R(\theta)
\]

\[
\nabla \Phi(\theta, Y) = -\frac{1}{T} A' \text{diag}(1 - \frac{r_n}{p_n(\theta)})
\]

where $A = \{a_{nj}\}$ is a large sparse matrix, and $\text{diag}(v_n)$ denotes a $N \times N$ diagonal matrix with elements $v_1, \ldots, v_N$ along the diagonal. For simplicity we focus on the case where $r_n = 0$, in which case $q_n(\theta) = p_n(\theta)$ and the above expressions simplify to

\[
\nabla^2 \Phi(\theta, Y) = A' \text{diag}(p_n(\theta)) A + \beta R(\theta)
\]

\[
\nabla \Phi(\theta, Y) = -\frac{1}{T} A'
\]

It follows from the assumption that the measurements have independent Poisson distributions that $\text{Cov}(Y) = \text{diag}\{Y_n(\theta^{\text{true}})\}$. Substituting into (11) and simplifying yields the following approximation to the estimator covariance:

\[
\text{Cov}(\hat{\theta}) \approx \frac{1}{T} \left[ F(\hat{\theta}) + \beta R(\hat{\theta}) \right]^{-1} F(\theta^{\text{true}}) (F(\hat{\theta}) + \beta R(\hat{\theta}))^{-1}
\]  
(26)

where

\[
F(\theta) = A' \text{diag}(p_n(\theta)) A
\]  
(27)

$^5$Due to the $\frac{1}{T}$ term in (25), one can show that for a fixed $\beta$, as $T \to \infty$, the maximum penalized-likelihood estimate $\hat{\theta}$ will converge in probability to $\theta$, a biased estimate [1]. For asymptotically unbiased estimates, one must let $\beta \to 0$ at an appropriate rate as $T \to \infty$ [6].
with the scale factor $T$ set so that $\sum_n Y_n(\theta^{\text{true}})$ was 0.25, 1, 4, and 16 million counts. We set $\sigma_n = 0$ for simplicity. For each study, we generated 100 realizations of pseudorandom Poisson transmission measurements according to (23) and then reconstructed using the penalized-likelihood estimator described by (25) using a coordinate-ascent algorithm [23]. This algorithm enforced the nonnegativity constraint $\theta \geq 0$. For simplicity, we used the function $\phi(x) = x^2/2$ for the penalty in (18). We also reconstructed attenuation maps using the conventional FBP algorithm at a matched resolution. The FBP images served as the initial estimate for the iterative algorithm.

We computed the sample standard deviations of the estimates for the center pixel from these simulations, as well as the approximate predicted variance given by (26). Fig. 2 shows the results as well as the (much inferior) performance of the conventional FBP method. The predicted variance agrees very well with the actual estimator performance, even for measured counts lower than are clinically relevant (20% error standard deviations would be clinically unacceptable). Therefore, for clinically relevant SNR's, the variance approximation given by (26) can be used to predict estimator performance reliably. For the simulation with 250K counts, the approximation agreed within 7% of the empirical results. For the simulations with more than 1M counts, the difference was smaller than 1%. Note the asymptotic property: better agreement between simulations and predictions for higher SNR.

Many authors have reported that the zeroth-order mean approximation (13) is reasonably accurate for maximum-likelihood estimators [14], [15], [20]; we have found similar results for penalized-likelihood estimators such as (25). This is fortuitous since the second-order expressions for mean are considerably more expensive to compute since $p = 128 \cdot 64$ and $N = 192 \cdot 96$ are very large in this example.) Fig. 3 displays a representative cross section through the mean predicted by (13) and the empirical sample mean computed from the 1M count simulations. The predicted mean agrees very closely with the sample mean. These results demonstrate that the mean and variance approximations (13) and (11) are useful for predicting penalized-likelihood estimator performance in transmission tomography.

V. POST-ESTIMATION PLUG-IN VARIANCE APPROXIMATION

The approximation (11) for the estimator covariance depends on both $\theta$ and $\text{Cov}\{Y\}$, so as written its primary use will be in computer simulations where $\theta$ and $\text{Cov}\{Y\}$ are known. Sometimes one would like to be able to obtain an approximate estimate of estimator variability from a single noisy measurement (such as real data), for which $\theta^{\text{true}}$ is unknown, and $\text{Cov}\{Y\}$ may also be unknown. In some problems this can be done using a “plug-in” estimate in which we substitute the estimate $\hat{\theta}$ in for $\theta$ in (11). The effectiveness of this approach will undoubtably be application-dependent, so in this section we focus on the specific problem of transmission tomography.

Using the transmission tomography model given in the previous section, assume we have a single noisy measurement realization $Y$ and a penalized-likelihood estimate $\hat{\theta}$ computed by maximizing the objective function (25). If we knew $\theta$ and $\theta^{\text{true}}$, then we could use (26) to approximate the covariance of $\hat{\theta}$. If we only have $\hat{\theta}$, then in light of the form of the covariance approximation given by (26), a natural approach to estimating the covariance would be to simply plug in $\hat{\theta}$ for $\theta$ and $\theta^{\text{true}}$ in (26)

$$\text{Cov}\{\hat{\theta}\} = \frac{1}{T}[F(\theta) + \beta R(\hat{\theta})]^{-1}F(\hat{\theta})[F(\hat{\theta}) + \beta R(\hat{\theta})]^{-1}$$

from which one can compute estimates of the variance of individual pixels or region-of-interest values using the same technique as in (12).

At first it may seem unlikely that such a simplistic approach would yield reliable estimates of variability. However, note that in definition (27) of $F(\theta)$, the only dependence on $\theta$ is
through its projections $p_n(\theta)$. In tomography, the projection operation is a smoothing operation, i.e., high spatial-frequency details are attenuated (hence the need for a ramp filter in linear reconstruction methods). Therefore, if the low and middle spatial frequencies of $\tilde{\theta}$ agree reasonably well with $\theta$ and $\theta^{\text{true}}$, then projections $p_n(\theta)$, $p_n(\tilde{\theta})$, and $p_n(\theta^{\text{true}})$ will be very similar. Furthermore, the dependence on the $p_n$ terms in (26) is through a diagonal matrix that is sandwiched between the $A'$ and $A$ matrices—which induce further smoothing.

To evaluate the reliability of this post-reconstruction plug-in estimate of variance, we used each of the 100 realizations described in the previous section to obtain a post-reconstruction estimate of the variance of estimate of the center pixel of the object shown in Fig. 1. If $\tilde{\theta}^{(m)}$ denotes the $m$th realization $(m = 1, \ldots, 100)$, then the $m$th estimate of the standard deviation of $\tilde{\theta}_j$ is

$$
\tilde{\sigma}_j^{(m)} = \left( \langle e^2 \rangle' \text{Cov}(\tilde{\theta}) e \right)^{1/2} = \left[ \langle e^2 \rangle' \left[ F(\hat{\theta}) + \beta R(\hat{\theta}) \right]^{-1} F(\hat{\theta}) \left[ F(\hat{\theta}) + \beta R(\hat{\theta}) \right]^{-1} e \right]^{1/2}.
$$

(28)

Histories of the standard deviation estimates $\{\tilde{\sigma}_j^{(m)}\}_{m=1}^{100}$ are shown in Figs. 4 and 5 for the 250K and 1M count simulations, respectively. The actual sample standard deviations for the two cases were $1.74 \cdot 10^{-3}$ and $9.30 \cdot 10^{-4}$, respectively. For the 250K count simulations, each of the 100 estimates was within 8% of the actual sample standard deviation. For the 1M count simulations, each of the 100 estimates was within 0.5% of the actual sample standard deviation. These are remarkably accurate estimates of variability, and clearly demonstrate the feasibility of estimating the variability of penalized-likelihood estimators even from single noisy measurements. One important application of such measures of variability would be in computing weighted estimates of kinetic parameters from dynamic PET scans [27].

VI. EXAMPLE: EMISSION TOMOGRAPHY

In this section we examine the accuracy of both the mean and the variance approximations for the problem of emission tomography. Our description of the problem is brief, for more details see [21], [28].

In emission tomography the parameter $\theta_j$ denotes the radionuclide concentration in the $j$th pixel. The emission measurements have independent Poisson distributions, and we assume the mean of $Y_j$ is

$$
Y_j(\theta) = T p_n(\theta)
$$

where $p_n(\theta) = \sum_j a_{nj} \theta_j + r_n$

(29)

where the $a_{nj}$ are proportional to the probability that an emission in voxel $j$ is detected by the $n$th detector pair, $\{r_n\}$ denotes additive background events such as random coincidences, and $T$ denotes the scan duration. These nonnegative factors are all assumed known. The log-likelihood for emission tomography has the same form as (24), but with definition (29) for $Y_j(\theta)$. We again focus on penalized-likelihood objective functions of the form (25).

Due to the nonnegativity constraints, the nonquadratic penalty (see below), and the nonquadratic form of the log-likelihood, this problem also provides a stringent test of the accuracy of our moment approximations.
A. Covariance Approximation

Approximating the variance of certain pixels of interest using (12) requires the following partial derivatives:

\[
\frac{\partial}{\partial \theta_j} L(\theta, Y) = T \sum_n a_{nj} \left( \frac{Y_n}{\bar{Y}_n(\theta)} - 1 \right)
\]

\[
\frac{\partial^2}{\partial \theta_j \partial \theta_k} L(\theta, Y) = T \sum_n a_{nj} a_{nk} \frac{Y_n}{\bar{Y}_n(\theta)}
\]

\[
\frac{\partial^2}{\partial \theta_j \partial \bar{Y}_n} L(\theta, Y) = a_{nj} / \bar{Y}_n(\theta).
\]

Combining the above expressions in matrix form with the expressions for the partials of \( R \) given in Section III

\[
-\nabla^{20} \Phi(\theta, Y) = A' \text{diag} \left\{ \frac{Y_n / T}{\bar{P}_n(\theta)} \right\} A + \beta R(\theta)
\]

\[
-\nabla^{11} \Phi(\theta, Y) = -\frac{1}{T} A' \text{diag} \left\{ \frac{1}{\bar{P}_n(\theta)} \right\}.
\]

Thus

\[
-\nabla^{20} \Phi(\hat{\theta}, \bar{Y}) = A' \text{diag} \left\{ \frac{P_n(\theta^{\text{true}})}{\bar{P}_n(\theta)} \right\} A + \beta R(\hat{\theta})
\]

\[
-\nabla^{11} \Phi(\hat{\theta}, \bar{Y}) = -\frac{1}{T} A' \text{diag} \left\{ \frac{1}{P_n(\theta)} \right\}.
\]

It follows from the assumption that the measurements have independent Poisson distributions that \( \text{Cov}\{Y\} = \text{diag}\{Y_n(\theta^{\text{true}})\} \). Substituting into (11) and simplifying yields the following approximation to the estimator covariance:

\[
\text{Cov}(\hat{\theta}) \approx \frac{1}{T} [F + \beta R(\hat{\theta})]^{-1} F [F + \beta R(\hat{\theta})]^{-1}
\]  \( \quad \text{(30)} \)

where

\[
F = A' \text{diag} \left\{ \frac{P_n(\theta^{\text{true}})}{\bar{P}_n(\theta)} \right\} A.
\]  \( \quad \text{(31)} \)

We compute the approximate variance of \( \hat{\theta}_j \) using a recipe similar to that given in Section IV.

B. Empirical Results

To assess the accuracy of approximation (30), we performed numerical simulations using the synthetic brain image shown in Fig. 6 as \( \theta^{\text{true}} \), with radioisotope concentrations 4 and 1 (arbitrary units) in gray and white matter, respectively. The image was a 112 by 128 array of 2 mm pixels. We simulated a PET emission scan with 80 radial bins and 110 angles uniformly spaced over 180°. The \( a_{nj} \) factors correspond to 6 mm wide strip integrals on 3 mm center-to-center spacing, modified by pseudorandom log-normal variates with a standard deviation of 0.3 to account for detector efficiency variations, and by head attenuation factors. Four studies were performed, with the scale factor \( T \) set so that \( \sum_n Y_n(\theta^{\text{true}}) \) was 0.2, 0.8, 3.2, and 12.8 million counts. The \( r_n \) factors were set to a uniform value corresponding to 10% random coincidences. For each study, 100 realizations of pseudorandom Poisson transmission measurements were generated according to (29) and then reconstructed using a space-alternating generalized EM algorithm [28], which enforces the nonnegativity constraint \( \hat{\theta} \geq 0 \). FBP images served as the initial estimate for the iterative algorithm.

For the penalty function \( \phi \) we studied two cases: the simple quadratic case \( \phi(x) = x^2/2 \), as well as a nonquadratic penalty: the third entry in [29, Table III]

\[
\phi(x) = \delta^2 [\|x\|/\delta - \log(1 + |x|/\delta)]
\]

with \( \delta = 1 \). This nonquadratic penalty blurs edges less than the quadratic penalty.

We computed the sample standard deviations of the estimates, as well as the approximate predicted variance given by (26) for two pixels: one at the center and one at the right edge of the left thalamus (oval shaped region near image center).

The results for the quadratic penalty are shown in Figs. 7 and 8. The trends are similar to those reported for transmission tomography: good agreement between the empirical standard
deviations and the analytical predictions, with improving accuracy with increasing counts. Note that for the quadratic penalty, pixels at the center and edge of the thalamus have similar variances.

The results for the nonquadratic penalty are shown in Figs. 9 and 10. For the pixel at the edge of the thalamus, the predicted and empirical variances agree well. But for the pixel at the center of the thalamus, the empirical variance was significantly higher than the predicted value for the 0.8M count case. Further work is therefore needed for nonquadratic penalties. Note that the edge pixel had higher variance than the center pixel with the nonquadratic penalty. The importance of this nonuniformity also needs investigation. Overall though, as in the transmission case we conclude that the variance approximation (11), (30) gives reasonably accurate predictions of estimator performance, with better agreement at higher SNR.

We also investigated the post-estimation plug-in approach described in Section V for the 0.8M count emission case. The plug-in estimates of standard deviation for the two pixels considered were all within 1% of the predicted values for the standard deviation. Thus plugging in $\hat{\theta}$ to (30) yields essentially the same value as one gets by using $\hat{\theta}$ and $\theta^{\text{true}}$. Therefore it appears that the intrinsic error in the approximation (30) is more significant than the differences between $\hat{\theta}$ and $\theta^{\text{true}}$. Practically, this suggests that if one can establish by simulation that the approximation error is small for measurements with more than a certain number of counts from a given tomograph, then one can use the plug-in approximation with such measurements and have confidence in the accuracy of the results even though $\theta^{\text{true}}$ is unknown.

As illustrated by Fig. 11, the zeroth-order mean approximation (13) again compares closely with the empirical sample mean for this likelihood-based estimator. However, the next subsection demonstrates that this accuracy does not apply to
the very nonlinear data-weighted least squares estimator for emission tomography.

C. Mean: Second Order

This subsection illustrates an application of the second-order approximation for estimator mean given by (16). In the routine practice of PET and SPECT, images are reconstructed using nonstatistical Fourier methods [30]. Often one can obtain more accurate images using likelihood-based methods. Since there is no closed-form expression for Poisson likelihood-based estimates, one must resort to iterative algorithms, many of which converge very slowly. Therefore, some investigators have replaced the log-likelihood objective with a weighted least squares or quadratic objective for which there are iterative algorithms that converge faster (e.g. [24], [25], [31], [32]). Unfortunately, in the context of transmission tomography, quadratic objectives lead to estimation bias for low-count measurements [23]. To determine whether a similar undesirable bias exists for the quadratic approximation in the emission case, we now use the analytical expression (16) for estimator mean.

The log-likelihood is nonquadratic, and the idea of using quadratic approximations to the log-likelihood has been studied extensively. Bouman and Sauer have nicely analyzed the approximations using a second-order Taylor expansion. Following [24], [25], the quadratic approximation to the log-likelihood \( \Phi_L(\theta, Y) = L(\theta, Y) \) is

\[
\Phi_Q(\theta, Y) = -\frac{1}{2} \sum_{n} \frac{1}{Y_n} (Y_n - Y_n(\theta))^2.
\]

The objective functions \( \Phi_L \) and \( \Phi_Q \) each implicitly define a nonlinear estimator. Even when \( p = 1 \), there is no closed-form solution for the maximum-likelihood estimate, except in the special case when \( r_n/a_n \) is a constant independent of \( n \).

For large images, the computation required for solving (16) appears prohibitive. Therefore, we consider a highly simplified version of emission tomography, where the unknown is a scalar parameter \( p = 1 \). This simplified problem nevertheless provides insight into the estimator bias without the undue notation of the multiparameter case. In Table I we derive the partial derivatives necessary for evaluating (16) for each objective (for \( p = 1 \)). In this table \( F_\theta \) denotes the Fisher information for estimating \( \theta \) from \( \{Y_n\} \)

\[
F_\theta = -E\{\nabla_\theta^2 \log f(Y; \theta)\} = \sum_n a_n^2/Y_n(\theta) = \sum_n a_n^2/\theta + r_n.
\]

The second and final two rows of Table I show three important points:

- For each objective, \( \nabla^2 \Phi(\theta, \bar{Y}(\theta)) = 0 \), so that \( \bar{\theta} = h(\bar{Y}(\theta)) = \theta \), i.e., the estimators work perfectly with noiseless data. Therefore, the 0th-order approximation (13) yields \( E(\bar{\theta}) = \theta \), which is inaccurate for the \( \Phi_Q \) estimator.
- The variances of the estimators are approximately equal.
- The maximum-likelihood estimate is unbiased to second order, whereas the quadratic estimate is biased.

Fig. 12 compares the bias predicted analytically using the approximation (16) with an empirically computed bias performed by numerical simulations. In these simulations we used \( \theta^\text{true} = 1, r_n = 0, a_n = 1 \), and \( N = 10 \), and varied \( T \) so that \( \bar{Y} = \sum Y_n(\theta^\text{true}) \) (average number of counts per detector) ranged from 2 to 100. The predicted and empirical results again agree very closely except when there are fewer than four average counts per detector. These results show that if the average counts per detector is below ten, then using the quadratic approximation to the Poisson log-likelihood can lead to biases exceeding 10%. In practice, the importance of this bias should be considered relative to other inaccuracies such as the approximations used in specifying \( a_n \). When the bias due to the quadratic approximation is significant, one can apply a hybrid Poisson/polynomial objective function similar to that proposed for transmission tomography [23]. In this approach, one uses the quadratic approximation for the high-count detectors, but the original log-likelihood for the low-count measurements, thereby retaining most of the computational advantage of the quadratic objective function without introducing bias [23].

VII. DISCUSSION

We have derived approximations for the mean and covariance of estimators that are defined as the maximum of some objective function. In the context of imaging applications with large numbers of unknown parameters, the variance approximation and the 0th-order mean approximation should be useful for predicting the performance of penalized-likelihood estimators. For applications with fewer parameters, one can also use the second-order mean approximation for improved accuracy.

In some applications one would like to perform estimation by maximizing an objective function subject to certain equality constraints. One can use methods similar to the derivation of the constrained Cramer–Rao lower bound [33], [34] to...
TABLE I
OBJECTIVE FUNCTIONS AND PARTIAL DERIVATIVES FOR SCALAR EMISSION TOMOGRAPHY PROBLEM WITH $\hat{Y}_n(\theta) = a_n \theta$

<table>
<thead>
<tr>
<th>Term</th>
<th>Likelihood</th>
<th>Quadratic</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\Phi(\theta, Y)$</td>
<td>$\sum_n Y_n \log \hat{Y}_n(\theta) - \hat{Y}_n(\theta)$</td>
<td>$-\frac{1}{2} \sum_n (Y_n - \hat{Y}_n(\theta))^2 / Y_n$</td>
</tr>
<tr>
<td>$\frac{\partial}{\partial \theta} \Phi(\theta, Y)$</td>
<td>$\sum_n a_n Y_n / \hat{Y}_n(\theta) - 1$</td>
<td>$\sum_n a_n (1 - \hat{Y}_n(\theta) / Y_n)$</td>
</tr>
<tr>
<td>$-\frac{\partial^2}{\partial \theta^2} \Phi(\theta, Y)$</td>
<td>$\sum_n a_n^2 Y_n / \hat{Y}_n(\theta)^2$</td>
<td>$\sum_n a_n^2 / Y_n$</td>
</tr>
<tr>
<td>$\frac{\partial^2}{\partial \theta^2} \Phi(\theta, Y)$</td>
<td>$2 \sum_n a_n^2 / Y_n^2$</td>
<td>$F_0$</td>
</tr>
<tr>
<td>$\frac{\partial^2}{\partial \theta^2} \Phi(\theta, Y)$</td>
<td>$a_n / \hat{Y}_n$</td>
<td>$F_0$</td>
</tr>
<tr>
<td>$\frac{\partial^2}{\partial \theta^2} \Phi(\theta, Y)$</td>
<td>$0$</td>
<td>$0$</td>
</tr>
<tr>
<td>$\frac{\partial^2}{\partial \theta^2} \Phi(\theta, Y)$</td>
<td>$-a_n^2 / \hat{Y}_n(\theta)^2$</td>
<td>$a_n^2 / Y_n^2$</td>
</tr>
<tr>
<td>$\frac{\sigma^2}{\eta_n} N(\tilde{Y})$</td>
<td>$1 / F_0$</td>
<td>$1 / F_0$</td>
</tr>
<tr>
<td>$\frac{\sigma^2}{\eta_n} N(\tilde{Y})$</td>
<td>$a_n / (\hat{Y}_n F_0)$</td>
<td>$a_n / (\hat{Y}_n F_0)$</td>
</tr>
<tr>
<td>$\frac{\sigma^2}{\eta_n} N(\tilde{Y})$</td>
<td>$\frac{1}{F_0^2} \sum_n a_n^2 / Y_n^2 - a_n / \hat{Y}_n$</td>
<td>$\frac{1}{F_0^2} \sum_n a_n^2 / Y_n^2 - a_n / \hat{Y}_n$</td>
</tr>
</tbody>
</table>

Our empirical results indicate that the accuracy of the proposed approximations improve with increasing SNR, which is consistent with the asymptotics discussed in the Appendix. If the SNR is too low, the approximation accuracy may be poor, but “how low is too low” will obviously be application-dependent. The approximations are also likely to overestimate the variance of pixels that are near zero when one enforces nonnegativity constraints. Thus these approximations do not eliminate the need for careful numerical simulations.

In our own work, thus far we have primarily used the approximations to determine useful values of the regularization parameter prior to performing simulations comparing various approaches (as in Section IV). In the future, we expect to evaluate the post-reconstruction estimate of region variability (Section V) for performing weighted estimates of kinetic parameters from dynamic PET emission scans [27]. Many PET scan protocols are indeed dynamic scans acquired for the purpose of extracting kinetic parameters; therefore, the ability to estimate region variability is essential. Since FBP is a linear reconstruction algorithm, it is straightforward to compute estimates of variability for Poisson emission measurements [27], [35]. If nonlinear penalized-likelihood methods are ever to replace FBP in the routine practice of PET, reliable estimates...
of variability (such as the plug-in method we have proposed) will be needed for a variety of purposes.

APPENDIX

This appendix synthesizes the asymptotic variance of $M$-estimates given by Serfling [1]. The results in Serfling are for a scalar parameter $\theta$, so we consider the scalar case below (see [36] for the multiparameter case). As in Section I, let $\Phi(\theta, Y)$ be the objective function that is to be maximized to find $\hat{\theta}$, and define

$$\psi(\theta, Y) = \frac{\partial}{\partial \theta} \Phi(\theta, Y).$$

Assume $Y$ has a probability distribution $F(y; \theta^{true})$, and let $\hat{\theta}$ be the value of $\theta$ that satisfies

$$\int \psi(\hat{\theta}, Y) dF(y; \theta^{true}) = 0. \quad (32)$$

Serfling [1] shows that $\hat{\theta}$ is asymptotically normal with mean $\theta$ and variance

$$\int \psi^2(\hat{\theta}, Y) dF(y; \theta^{true}) \left[ \int \psi^2(\theta, Y) dF(y; \theta^{true}) \right]^{-1}. \quad (33)$$

This asymptotic variance is somewhat inconvenient to use in imaging problems for the following reasons:

- The term $\hat{\theta}$ plays a role similar to our $\hat{\theta}$, but solving the integral equation (32) for $\hat{\theta}$ is in general more work than calculating $\hat{\theta}$ by maximizing $\Phi(\cdot, Y)$.
- Both $\hat{\theta}$ and the expression for the asymptotic variance depend on the entire measurement distribution $F(y; \theta^{true})$, whereas our approximation depends only on the mean and covariance of the measurements.

With some additional work, one can show that if $\psi(\theta, Y)$ is affine in $Y$, then $\hat{\theta}$ and $\hat{\theta}$ are equal, and (33) is equivalent to (11). Both Gaussian and Poisson measurements yield $\psi$ that are affine in $Y$ (cf (24)), so (11) is the asymptotic covariance in those cases, provided the penalty is data-independent. For data-dependent penalties [37] or for more complicated noise distributions, such as the Poisson/Gaussian model for CCD arrays [38], the covariance approximation given by (11) will probably be easier to implement than (33).

REFERENCES


**Jeffrey A. Fessler** (S’83–M’90) received the B.S.E.E. degree from Purdue University, West Lafayette, IN, in 1985, the M.S.E.E. degree from Stanford University, Stanford, CA, in 1986, and the M.S. degree in statistics, also from Stanford University, in 1989. From 1985 to 1988, he was a National Science Foundation Graduate Fellow at Stanford, where he received the Ph.D. degree in electrical engineering in 1990.

From 1991 to 1992, he was a Department of Energy Alexander Hollaender Post-Doctoral Fellow in the Division of Nuclear Medicine at the University of Michigan, Ann Arbor. From 1993 to 1995, he was an Assistant Professor in Nuclear Medicine and the Bioengineering Program of the University of Michigan. Since 1995, he has been with the Department of Electrical Engineering and Computer Science of the University of Michigan. His research interests are in statistical aspects of medical imaging.