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

Jeffrey A. Fessler
Email: fessler@umich.edu, Phone: 734-763-1434

IEEE T-MI 5(3):493-506, Mar. 1996

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 a unknown real parameter vector that is to be estimated from a measurement vector $Y = [Y_1, \ldots, Y_N]' \in \mathbb{R}^N$, where $'$ 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:

$$\hat{\theta} = \arg\max_{\theta} \Phi(\theta, Y). \qquad (1)$$

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 *implicitly* defines $\hat{\theta}$ as a function of Y. Statisticians refer to (1) as an *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.

This work was supported in part by DOE grant DE-FG02-87ER60561 and NIH grants CA-60711 and CA-54362.

Because of the partial derivatives used in the derivation, our approximations are restricted to problems where θ 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 $\hat{\theta}$, 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 $\hat{\theta}$. Although this constraint is often important in *unpenalized* estimation methods, our primary interest is in objective functions $\Phi(\theta, Y)$ 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 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 $\Phi(\cdot, Y)$ has a unique global maximum $\hat{\theta} \in \Theta$ for any measurement Y, so that $\hat{\theta}$ 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 $\Phi(\cdot, Y)$:

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

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

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

$$0 = \frac{\partial}{\partial \theta_j} \Phi(\theta, Y) \bigg|_{\theta=h(Y)}, \ j = 1, \ldots, p. \tag{3}$$

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

$$0 = \frac{\partial}{\partial \theta_j} \Phi(h(Y), Y), \ j = 1, \ldots, p, \tag{4}$$

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 $\frac{\partial}{\partial Y_n}$ 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 $\hat{\theta}$. 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 $\bar{Y}_n$ denotes the mean of Y_n, then

$$\begin{aligned} h(Y) &\approx h(\bar{Y}) + \sum_n \frac{\partial}{\partial Y_n} h(\bar{Y})(Y_n - \bar{Y}_n) \\ &+ \frac{1}{2} \sum_n \sum_m \frac{\partial^2}{\partial Y_n \partial Y_m} h(\bar{Y})(Y_n - \bar{Y}_n)(Y_m - \bar{Y}_m) \end{aligned} \tag{5}$$

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:

$$h(Y) \approx h(\bar{Y}) + \nabla h(\bar{Y})(Y - \bar{Y}), \tag{6}$$

where $\nabla = [\frac{\partial}{\partial Y_1} \; \dots \; \frac{\partial}{\partial Y_N}]$ denotes the (row) gradient operator. Taking the covariance[2] of both sides yields the following well-known approximation [17, p. 426]:

$$\mathrm{Cov}\{\hat{\theta}\} = \mathrm{Cov}\{h(Y)\} \approx \nabla h(\bar{Y}) \, \mathrm{Cov}\{Y\} \, [\nabla h(\bar{Y})]'. \qquad (7)$$

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[3] yields:

$$0 = \sum_k \frac{\partial^2}{\partial\theta_j \partial\theta_k} \Phi(h(Y), Y) \, \frac{\partial}{\partial Y_n} h_k(Y) + \frac{\partial^2}{\partial\theta_j \partial Y_n} \Phi(h(Y), Y), \qquad (8)$$

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

$$0 = \nabla^{20} \Phi(h(\bar{Y}), \bar{Y}) \, \nabla h(\bar{Y}) + \nabla^{11} \Phi(h(\bar{Y}), \bar{Y}), \qquad (9)$$

where the (j, k)th element of the $p \times p$ operator ∇^{20} is $\frac{\partial^2}{\partial\theta_j \partial\theta_k}$, and the (j, n)th element of the $p \times N$ operator ∇^{11} is $\frac{\partial^2}{\partial\theta_j \partial Y_n}$. To proceed, we now assume that the symmetric matrix $-\nabla^{20} \Phi(h(\bar{Y}), \bar{Y})$ is also positive definite[4], so we can solve for $\nabla h(\bar{Y})$ by rearranging:

$$\nabla h(\bar{Y}) = [-\nabla^{20} \Phi(h(\bar{Y}), \bar{Y})]^{-1} \, \nabla^{11} \Phi(h(\bar{Y}), \bar{Y}). \qquad (10)$$

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

$$\mathrm{Cov}\{\hat{\theta}\} \approx [-\nabla^{20} \Phi(\check{\theta}, \bar{Y})]^{-1} \, \nabla^{11} \Phi(\check{\theta}, \bar{Y}) \, \mathrm{Cov}\{Y\} \cdot [\nabla^{11} \Phi(\check{\theta}, \bar{Y})]' \, [-\nabla^{20} \Phi(\check{\theta}, \bar{Y})]^{-1}. \qquad (11)$$

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 jth unit vector of length p, and define $u^j = [-\nabla^{20} \Phi(\check{\theta}, \bar{Y})]^{-1} e^j$. Note that one does not need to perform a $p \times p$ matrix inversion to compute u^j; one simply solves the equation $[-\nabla^{20} \Phi(\check{\theta}, \bar{Y})] u^j = e^j$, 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

$$\mathrm{Cov}\{\hat{\theta}_j, \hat{\theta}_k\} = (e^j)' \mathrm{Cov}\{\hat{\theta}\} e^k$$
$$\approx (u^j)' \, \nabla^{11} \Phi(\check{\theta}, \bar{Y}) \, \mathrm{Cov}\{Y\} \, [\nabla^{11} \Phi(\check{\theta}, \bar{Y})]' u^k, \qquad (12)$$

for $j, k = 1, \dots, p$. One can compute any portion of the covariance matrix of $\hat{\theta}$ by using (12) repeatedly for appropriate j and k. In general, computing $\mathrm{Var}\{\hat{\theta}_j\}$ 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[5] 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:

$$E\{\hat{\theta}\} = E\{h(Y)\} \approx h(\bar{Y}) = \check{\theta}. \qquad (13)$$

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 Φ, so (13) allows one to examine the effects of the penalty separately from the effects of noise. However, the approximation (13) is certainly not always adequate, as the example in Section VI illustrates. Therefore, 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 approximation for the mean of $h(Y)$:

$$E\{\hat{\theta}\} \approx h(\bar{Y}) + \frac{1}{2} \sum_n \sum_m \frac{\partial^2}{\partial Y_n \partial Y_m} h(\bar{Y}) \mathrm{Cov}\{Y_n, Y_m\}, \qquad (14)$$

where $\mathrm{Cov}\{Y_n, Y_m\} = E\{(Y_n - \bar{Y}_n)(Y_m - \bar{Y}_m)\}$ 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 Y_m, obtaining:

$$0 = \sum_k \left[\sum_l \frac{\partial^3}{\partial\theta_j \partial\theta_k \partial\theta_l} \Phi(h(Y), Y) \frac{\partial}{\partial Y_m} h_l(Y) \right.$$

[2] All expectations and covariances are taken with respect to the probability density of the random measurement Y. Typically one assumes this density is of the form $f(Y; \theta^{\mathrm{true}})$, where θ^{true} is the unknown parameter to be estimated using (1). However, our approximations do not *require* a parametric form for the measurement distribution; we need only that the covariance of the measurements be known (or can be estimated—see Section V).

[3] We restrict attention to objective functions $\Phi(\theta, Y)$ for which the partial derivatives we use exist.

[4] The assumption that $-\nabla^{20} \Phi(h(\bar{Y}), \bar{Y})$ is positive definite is much less restrictive than the usual assumption that $\Phi(\cdot, Y)$ is globally strictly concave for any measurement vector Y. We only require that $\Phi(h(\bar{Y}), \bar{Y})$ be locally strictly concave (near $\check{\theta}$) for noise-free data $\bar{Y}$.

[5] Note that (11) and (12) do depend on $\check{\theta} = h(\bar{Y})$. By the definition (3) of $h(Y)$, we see that $\check{\theta} = \arg\max_\theta \Phi(\theta, \bar{Y})$, so we compute $\check{\theta}$ by applying the estimation algorithm to the noise free data $\bar{Y}$.

$$+\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial Y_m}\Phi(h(Y),Y)\Bigg]\frac{\partial}{\partial Y_n}h_k(Y)$$

$$+\sum_k\frac{\partial^2}{\partial\theta_j\partial\theta_k}\Phi(h(Y),Y)\;\frac{\partial^2}{\partial Y_n\partial Y_m}h_k(Y)$$

$$+\sum_k\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial Y_n}\Phi(h(Y),Y)\frac{\partial}{\partial Y_m}h_k(Y)$$

$$+\frac{\partial^3}{\partial\theta_j\partial Y_n\partial Y_m}\Phi(h(Y),Y), \qquad (15)$$

for $j=1,\ldots,p,\; n=1,\ldots,N,\; m=1,\ldots,N$. One can substitute in $\check{\theta}=h(\bar{Y})$ and $Y=\bar{Y}$ in the above expression to obtain N^2 sets of p equations in the p unknowns $\{\frac{\partial^2}{\partial Y_n\partial Y_m}h_k(\bar{Y})\}_{k=1}^{p}$. 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 Φ. Note that (15) depends on the first partials $\frac{\partial}{\partial Y_n}h_k(Y)$, 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 Φ, described next.

C. Independent Measurements

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

$$E\{\hat{\theta}\}=E\{h(Y)\}\approx h(\bar{Y})+\frac{1}{2}\sum_n\frac{\partial^2}{\partial Y_n^2}h(\bar{Y})\mathrm{Var}\{Y_n\}. \quad (16)$$

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:

$$0=\sum_k\Bigg[\sum_l\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\theta_l}\Phi(h(Y),Y)\frac{\partial}{\partial Y_n}h_l(Y)$$

$$+2\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial Y_n}\Phi(h(Y),Y)\Bigg]\frac{\partial}{\partial Y_n}h_k(Y)$$

$$+\sum_k\frac{\partial^2}{\partial\theta_j\partial\theta_k}\Phi(h(Y),Y)\;\frac{\partial^2}{\partial Y_n^2}h_k(Y)$$

$$+\frac{\partial^3}{\partial\theta_j\partial Y_n^2}\Phi(h(Y),Y),\; j=1,\ldots,p,\; n=1,\ldots,N.$$

D. Scalar Parameter

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

$$0=\Bigg[\frac{\partial^3}{\partial\theta^3}\Phi(h(Y),Y)\frac{\partial}{\partial Y_m}h(Y)$$

$$+\frac{\partial^3}{\partial\theta^2\partial Y_m}\Phi(h(Y),Y)\Bigg]\frac{\partial}{\partial Y_n}h(Y)$$

$$+\frac{\partial^2}{\partial\theta^2}\Phi(h(Y),Y)\;\frac{\partial^2}{\partial Y_n\partial Y_m}h(Y)$$

$$+\frac{\partial^3}{\partial\theta^2\partial Y_n}\Phi(h(Y),Y)\frac{\partial}{\partial Y_m}h(Y)$$

$$+\frac{\partial^3}{\partial\theta\partial Y_n\partial Y_m}\Phi(h(Y),Y),\; n=1,\ldots,N,\; m=1,\ldots,N. \quad (17)$$

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

$$\frac{\partial^2}{\partial Y_n\partial Y_m}h(\bar{Y})=\left[-\frac{\partial^2}{\partial\theta^2}\Phi(\check{\theta},\bar{Y})\right]^{-1}$$

$$\Bigg(\Bigg[\frac{\partial^3}{\partial\theta^3}\Phi(\check{\theta},\bar{Y})\frac{\partial}{\partial Y_m}h(\bar{Y})+\frac{\partial^3}{\partial\theta^2\partial Y_m}\Phi(\check{\theta},\bar{Y})\Bigg]\frac{\partial}{\partial Y_n}h(\bar{Y})$$

$$+\frac{\partial^3}{\partial\theta^2\partial Y_n}\Phi(\check{\theta},\bar{Y})\frac{\partial}{\partial Y_m}h(\bar{Y})+\frac{\partial^3}{\partial\theta\partial Y_n\partial Y_m}\Phi(\check{\theta},\bar{Y})\Bigg).$$

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:

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

where $\mathbf{A}$ is a known $N\times p$ matrix. For such problems, the following regularized weighted least-squares objective function is often used for estimation:

$$\Phi(\theta,Y)=-\frac{1}{2}(Y-\mathbf{A}\theta)'\mathbf{\Pi}(Y-\mathbf{A}\theta)-\beta R(\theta),$$

where $\mathbf{\Pi}$ is a nonnegative definite weighting matrix and $R(\theta)$ is a roughness penalty of the form

$$R(\theta)=\sum_j\frac{1}{2}\sum_k w_{jk}\phi(\theta_j-\theta_k), \quad (18)$$

where $w_{jk}=1$ for horizontal and vertical neighbors, $w_{jk}=1/\sqrt{2}$ for diagonal neighbors, and is 0 otherwise. Note that

$$\nabla^{11}R(\theta)=0$$

and define

$$\mathbf{R}(\theta) = \nabla^2 R(\theta) \tag{19}$$

to be the matrix of second partials of $R(\theta)$. The (j,k)th element of $\mathbf{R}(\theta)$ is:

$$\begin{cases} \sum_{k'} w_{jk'}\ddot{\phi}(\theta_j - \theta_{k'}), & j = k \\ -w_{jk}\ddot{\phi}(\theta_j - \theta_k), & j \neq k \end{cases},$$

where $\ddot{\phi}$ denotes the second derivative of ϕ.

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 $\mathbf{\Pi A}$. In this case one can derive an explicit expression for the estimator:

$$\hat{\theta} = h(Y) = (\mathbf{A}'\mathbf{\Pi A} + \beta\mathbf{R})^{-1}\mathbf{A}'\mathbf{\Pi}Y, \tag{20}$$

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:

$$\begin{aligned} -\nabla^{20}\Phi &= \mathbf{A}'\mathbf{\Pi A} + \beta\mathbf{R} \\ \nabla^{11}\Phi &= -\mathbf{A}'\mathbf{\Pi} \\ \nabla^{30}\Phi &= \nabla^{21}\Phi = \nabla^{12}\Phi = 0, \end{aligned} \tag{21}$$

so substituting into (15), one finds that $\nabla^2 h(Y)$ = 0. Thus from (14):

$$E\{\hat{\theta}\} = h(\bar{Y}) = (\mathbf{A}'\mathbf{\Pi A} + \beta\mathbf{R})^{-1}\mathbf{A}'\mathbf{\Pi}\bar{Y},$$

which of course is exactly what one would get from (20). Substituting (21) into (11) yields the estimator covariance:

$$\mathrm{Cov}\{\hat{\theta}\} =$$

$$[\mathbf{A}'\mathbf{\Pi A} + \beta\mathbf{R}]^{-1}\mathbf{A}'\mathbf{\Pi}\mathrm{Cov}\{Y\}\mathbf{\Pi A}[\mathbf{A}'\mathbf{\Pi A} + \beta\mathbf{R}]^{-1},$$

which again agrees with (20). If the measurement covariance is known, then usually one chooses $\mathbf{\Pi}^{-1} = \mathrm{Cov}\{Y\}$, in which case

$$\mathrm{Cov}\{\hat{\theta}\} = (\mathbf{F} + \beta\mathbf{R})^{-1}\mathbf{F}(\mathbf{F} + \beta\mathbf{R})^{-1}, \tag{22}$$

where $\mathbf{F} = \mathbf{A}'\mathrm{Cov}\{Y\}^{-1}\mathbf{A}$ is the Fisher information for estimating θ from Y, when the noise has a normal distribution. The covariance approximations derived in the following sections are similar to (22).

Since our approximations for the mean and covariance are exact for quadratic objective functions, one might expect the approximation accuracy for a non-quadratic objective will depend on how far the objective deviates from being quadratic. Many objective functions are locally quadratic, so we expect that the approximation accuracy will depend on the signal to noise ratio (SNR) of the measurements. Indeed, from (5) it is clear that as the noise variance goes to zero, we will have $Y_n \to \bar{Y}_n$, so the Taylor approximation error will vanish. This asymptotic property is illustrated empirically in the next section.

IV. Example: Transmission Tomography

To illustrate the accuracy of the approximation for estimator covariance given by (11), in this section we consider the problem of tomographic reconstruction from Poisson distributed PET transmission data. Our description of the problem is brief, for more details see [21–23]. Since PET transmission scans are essentially measurements of nuisance parameters, one would like to use very short transmission scans. Since short scans have fewer counts (lower SNR), the conventional linear filtered backprojection (FBP) reconstruction method performs poorly. Statistical methods have the potential to significantly reduce the error variance, but since they are nonlinear, only empirical studies of estimator performance have been previously performed to our knowledge. Analytical expressions for the variance will help us determine (without exhaustive simulations) conditions under which statistical methods will outperform FBP.

In transmission tomography the parameter θ_j denotes the attenuation coefficient in the jth pixel. The transmission measurements have independent Poisson distributions, and we assume the mean of Y_n is:

$$\begin{aligned} \bar{Y}_n(\theta) &= Tp_n(\theta) \\ p_n(\theta) &= b_n e^{-\sum_j a_{nj}\theta_j} + r_n, \end{aligned} \tag{23}$$

where the a_{nj} factors denote the intersection length of the nth ray passing though the jth pixel, $\{b_n\}$ denote the rates of emissions from the transmission source, $\{r_n\}$ denote additive background events such as random coincidences, and T denotes the scan duration. These nonnegative factors are all assumed known. The log-likelihood is:

$$L(\theta, Y) = \sum_n Y_n \log \bar{Y}_n(\theta) - \bar{Y}_n(\theta), \tag{24}$$

neglecting constants independent of θ. Since tomography is ill-conditioned, rather than performing ordinary ML estimation, many investigators have used penalized-likelihood objective functions of the form[6]

$$\Phi(\theta, Y) = \frac{1}{T}L(\theta, Y) - \beta R(\theta), \tag{25}$$

where the roughness penalty R was defined in (18).

Due to the nonlinearity of (23) and the non-quadratic 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),

[6] Due to the $\frac{1}{T}$ term in (25), one can show that for a fixed β, as $T \to \infty$, the maximum penalized-likelihood estimate $\hat{\theta}$ will converge in probability to $\check{\theta}$, a biased estimate [1]. For asymptotically unbiased estimates, one must let $\beta \to 0$ at an appropriate rate as $T \to \infty$ [6].

which requires the following partial derivatives:

$$\begin{aligned}
\frac{\partial}{\partial \theta_j} L(\theta, Y) &= T \sum_n a_{nj} \left(1 - \frac{Y_n}{\bar{Y}_n(\theta)}\right) (p_n(\theta) - r_n) \\
-\frac{\partial^2}{\partial \theta_j \partial \theta_k} L(\theta, Y) &= T \sum_n a_{nj} a_{nk} q_n(\theta) \\
q_n(\theta) &= \left(1 - \frac{r_n Y_n / T}{p_n^2(\theta)}\right) b_n e^{-\sum_j a_{nj}\theta_j} \\
\frac{\partial^2}{\partial \theta_j \partial Y_n} L(\theta, Y) &= -a_{nj} \left(1 - \frac{r_n}{p_n(\theta)}\right).
\end{aligned}$$

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

$$\begin{aligned}
-\nabla^{20}\Phi(\theta, Y) &= \mathbf{A}' \mathrm{diag}\left\{q_n(\theta)\right\} \mathbf{A} + \beta \mathbf{R}(\theta) \\
\nabla^{11}\Phi(\theta, Y) &= -\frac{1}{T} \mathbf{A}' \mathrm{diag}\left\{1 - \frac{r_n}{p_n(\theta)}\right\},
\end{aligned}$$

where $\mathbf{A} = \{a_{nj}\}$ is a large sparse matrix, and $\mathrm{diag}\{v_n\}$ 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

$$\begin{aligned}
-\nabla^{20}\Phi(\theta, Y) &= \mathbf{A}' \mathrm{diag}\left\{p_n(\theta)\right\} \mathbf{A} + \beta \mathbf{R}(\theta) \\
\nabla^{11}\Phi(\theta, Y) &= -\frac{1}{T} \mathbf{A}'.
\end{aligned}$$

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

$$\mathrm{Cov}\{\hat{\theta}\} \approx \frac{1}{T} [\mathbf{F}(\check{\theta}) + \beta \mathbf{R}(\check{\theta})]^{-1} \mathbf{F}(\theta^{\mathrm{true}}) [\mathbf{F}(\check{\theta}) + \beta \mathbf{R}(\check{\theta})]^{-1}, \tag{26}$$

where

$$\mathbf{F}(\theta) = \mathbf{A}' \mathrm{diag}\left\{p_n(\theta)\right\} \mathbf{A} \tag{27}$$

is $1/T$ times the Fisher information for estimating θ from Y. Note the similarity to (22).

We compute the approximate variance of $\hat{\theta}_j$ by using the following recipe.

- Compute $\check{\theta} = \arg\max_\theta \Phi(\theta, \bar{Y})$ by applying to noise-free data $\bar{Y}$ a maximization algorithm such as the fast converging coordinate-ascent algorithm of Bouman and Sauer [24, 25].

- Forward project $\check{\theta}$ to compute $p_n(\check{\theta}) = \sum_j a_{nj} \check{\theta}_j + r_n$. Likewise for $p_n(\theta^{\mathrm{true}})$.

- Pick a pixel j of interest and solve the equation $[\mathbf{A}' \mathrm{diag}\left\{p_n(\check{\theta})\right\} \mathbf{A} + \beta \mathbf{R}(\check{\theta})] u^j = e^j$ for u^j using a fast iterative method such as preconditioned conjugate gradients [26] or Gauss-Siedel [19].

- Compute $\frac{1}{T} (u^j)' \mathbf{A}' \mathrm{diag}\left\{p_n(\theta^{\mathrm{true}})\right\} \mathbf{A} u^j$ by first forward projecting u^j to compute $v = \mathbf{A} u^j$, and then summing:

$$\mathrm{Var}\{\hat{\theta}_j\} \approx \frac{1}{T} \sum_n v_n^2 p_n(\theta^{\mathrm{true}}).$$

The overall computational requirements for this recipe are roughly equivalent to two maximizations of Φ. Thus, if one only needs the approximate variance for a few pixels of interest, it is more efficient to use the above technique than to perform numerical simulations that require dozens of maximizations of Φ.

B. Empirical Results

To assess the accuracy of approximation (26), we performed numerical simulations using the synthetic attenuation map shown in Fig. 1 as θ^{true}. This image represents a human thorax cross-section with linear attenuation coefficients 0.0165mm^{-1}, 0.0096mm^{-1}, and 0.0025mm^{-1}, for bone, soft tissue, and lungs respectively. The image was a 128 by 64 array of 4.5mm pixels. We simulated a PET transmission scan with 192 radial bins and 96 angles uniformly spaced over 180°. The a_{nj} factors corresponded to 6mm wide strip integrals with 3mm center-to-center spacing. (This is an approximation to the ideal line integral that accounts for finite detector width.) We generated the b_n factors using pseudo-random log-normal variates with a standard deviation of 0.3 to account for detector efficiency variations. We performed four studies with the scale factor T set so that $\sum_n \bar{Y}_n(\theta^{\mathrm{true}})$ was 0.25, 1, 4, and 16 million counts. We set $r_n = 0$ for simplicity. For each study, we generated 100 realizations of pseudo-random 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 $\hat{\theta} \geq 0$. 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 SNRs, 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 0th-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 2nd-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.) Figure 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 $\check{\theta}$ and $\mathrm{Cov}\{Y\}$, so as written its primary use will be in computer simulations where $\check{\theta}$ and $\mathrm{Cov}\{Y\}$ 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 θ^{true} is unknown, and $\mathrm{Cov}\{Y\}$ 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 $\check{\theta}$ 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 $\check{\theta}$ and θ^{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 $\check{\theta}$ and θ^{true} in (26):

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

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 the definition (27) of $\mathbf{F}(\theta)$, the *only* dependence on θ 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 $\hat{\theta}$ agree reasonably well with $\check{\theta}$ and θ^{true}, then the projections $p_n(\hat{\theta})$, $p_n(\check{\theta})$, and $p_n(\theta^{\mathrm{true}})$ will be very similar. Furthermore, the dependence on the p_n terms in (26) is through a diagonal matrix that is sandwiched between the $\mathbf{A}'$ and $\mathbf{A}$ 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 $\hat{\theta}^{(m)}$ denotes the mth realization $(m = 1, \ldots, 100)$, then the mth estimate of the standard deviation of $\hat{\theta}_j$ is:

$$\hat{\sigma}_j^{(m)} = \left[(e^j)'\widehat{\mathrm{Cov}\{\hat{\theta}\}}e^j\right]^{1/2}$$
$$= \left[(e^j)'\frac{1}{T}[\mathbf{F}(\hat{\theta}) + \beta\mathbf{R}(\hat{\theta})]^{-1}\mathbf{F}(\hat{\theta})[\mathbf{F}(\hat{\theta}) + \beta\mathbf{R}(\hat{\theta})]^{-1}e^j\right]^{1/2}. \tag{28}$$

Histograms of the standard deviation estimates $\left\{\hat{\sigma}_j^{(m)}\right\}_{m=1}^{100}$ 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 θ_j denotes the radionuclide concentration in the jth pixel. The emission measurements have independent Poisson distributions, and we assume the mean of Y_n is:

$$\begin{aligned} \bar{Y}_n(\theta) &= Tp_n(\theta) \\ p_n(\theta) &= \sum_j a_{nj}\theta_j + r_n, \end{aligned} \tag{29}$$

where the a_{nj} are proportional to the probability that an emission in voxel j is detected by the nth 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 $\bar{Y}_n(\theta)$. 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:

$$\begin{aligned} \frac{\partial}{\partial\theta_j}L(\theta, Y) &= T\sum_n a_{nj}\left(\frac{Y_n}{\bar{Y}_n(\theta)} - 1\right) \\ -\frac{\partial^2}{\partial\theta_j\partial\theta_k}L(\theta, Y) &= T\sum_n a_{nj}a_{nk}\frac{Y_n/T}{p_n^2(\theta)} \end{aligned}$$

$$\frac{\partial^2}{\partial\theta_j \partial Y_n} L(\theta, Y) \quad = \quad a_{nj}/p_n(\theta).$$

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

$$\begin{aligned} -\nabla^{20}\Phi(\theta, Y) &= \mathbf{A}'\mathrm{diag}\left\{\frac{Y_n/T}{p_n^2(\theta)}\right\}\mathbf{A} + \beta\mathbf{R}(\theta) \\ \nabla^{11}\Phi(\theta, Y) &= -\frac{1}{T}\mathbf{A}'\mathrm{diag}\left\{\frac{1}{p_n(\theta)}\right\}. \end{aligned}$$

Thus

$$\begin{aligned} -\nabla^{20}\Phi(\check{\theta}, \bar{Y}) &= \mathbf{A}'\mathrm{diag}\left\{\frac{p_n(\theta^{\mathrm{true}})}{p_n^2(\check{\theta})}\right\}\mathbf{A} + \beta\mathbf{R}(\check{\theta}) \\ \nabla^{11}\Phi(\check{\theta}, \bar{Y}) &= -\frac{1}{T}\mathbf{A}'\mathrm{diag}\left\{\frac{1}{p_n(\check{\theta})}\right\}. \end{aligned}$$

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

$$\mathrm{Cov}\{\hat{\theta}\} \approx \frac{1}{T}[\mathbf{F} + \beta\mathbf{R}(\check{\theta})]^{-1}\mathbf{F}[\mathbf{F} + \beta\mathbf{R}(\check{\theta})]^{-1}, \qquad (30)$$

where

$$\mathbf{F} = \mathbf{A}'\mathrm{diag}\left\{\frac{p_n(\theta^{\mathrm{true}})}{p_n^2(\check{\theta})}\right\}\mathbf{A}. \qquad (31)$$

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 θ^{true}, with radioisotope concentrations 4 and 1 (arbitrary units) in gray and white matter respectively. The image was a 112 by 128 array of 2mm pixels. We simulated a PET emission scan with 80 radial bins and 110 angles uniformly spaced over 180°. The a_{nj} factors correspond to 6mm wide strip integrals on 3mm center-to-center spacing, modified by pseudo-random 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 $\sum_n \bar{Y}_n(\theta^{\mathrm{true}})$ 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 pseudo-random 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 ϕ we studied two cases: the simple quadratic case $\phi(x) = x^2/2$, as well as a nonquadratic penalty: the third entry in Table III of [29]:

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

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 $\check{\theta}$ and θ^{true}. Thus it appears that the intrinsic error in the approximation (30) is more significant than the differences between $\hat{\theta}$ and θ^{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 θ^{true} is unknown.

As illustrated by Fig. 11, the 0th-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: 2nd 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 non-statistical 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 non-quadratic, 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

$$\Phi_Q(\theta, Y) = -\frac{1}{2} \sum_{n \,:\, Y_n > 0} \frac{1}{Y_n} (Y_n - \bar{Y}_n(\theta))^2.$$

The objective functions Φ_L and Φ_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 multi-parameter case. In Table 1 we derive the partial derivatives necessary for evaluating (16) for each objective (for $p = 1$). In this table F_θ denotes the Fisher information for estimating θ from $\{Y_n\}$:

$$F_\theta = -E\{\nabla^2_\theta \log f(Y, \theta)\} = \sum_n a_n^2 / \bar{Y}_n(\theta) = \sum_n \frac{a_n^2}{a_n \theta + r_n}.$$

The second and final two rows of Table 1 show three important points:

- For each objective, $\nabla^{10} \Phi(\theta, \bar{Y}(\theta)) = 0$, so that $\check{\theta} = h(\bar{Y}(\theta)) = \theta$, i.e. the estimators work perfectly with noiseless data. Therefore the 0th-order approximation (13) yields $E\{\hat{\theta}\} = \theta$, which is inaccurate for the Φ_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.

Figure 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 $\frac{1}{N} \sum_n \bar{Y}_n(\theta^{\text{true}})$ (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 4 average counts per detector. These results show that if the average counts per detector is below 10, 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 generalize the covariance approximation (11) to include the reduction in variance that results from including constraints.

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.

VIII. Appendix

This appendix synopsizes the asymptotic variance of M-estimates given by Serfling [1]. The results in Serfling are for a scalar parameter θ, 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 $F(y; \theta^{\mathrm{true}})$, and let $\bar{\theta}$ be the value of θ that satisfies:

$$\int \psi(\theta, y) \; dF(y; \theta^{\mathrm{true}}) = 0. \qquad (32)$$

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

$$\frac{\int \psi^2(\bar{\theta}, y) \; dF(y; \theta^{\mathrm{true}})}{\left[\frac{\partial}{\partial \theta} \int \psi^2(\theta, y) \; dF(y; \theta^{\mathrm{true}}) \big|_{\theta = \bar{\theta}} \right]^2}. \qquad (33)$$

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

- The term $\bar{\theta}$ plays a role similar to our $\check{\theta}$, but solving the integral equation (32) for $\bar{\theta}$ is in general more work than calculating $\check{\theta}$ by maximizing $\Phi(\cdot, \bar{Y})$.
- Both $\bar{\theta}$ and the expression for the asymptotic variance depend on the entire measurement distribution $F(y; \theta^{\mathrm{true}})$, 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 $\bar{\theta}$ and $\check{\theta}$ are equal, and (33) is equivalent to (11). Both Gaussian and Poisson measurements yield ψ 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

[1] R. J. Serfling. *Approximation theorems of mathematical statistics*. Wiley, New York, 1980.

[2] A. O. Hero. A Cramer-Rao type lower bound for essentially unbiased parameter estimation. Technical Report 890, Lincoln Laboratory, MIT, ?, January 1992.

[3] J. A. Fessler and A. O. Hero. Cramer-Rao lower bounds for biased image reconstruction. In *Proc. Midwest Symposium on Circuits and Systems*, volume 1, pages 253–256, 1993.

[4] J. A. Fessler. Moments of implicitly defined estimators (e.g. ML and MAP): applications to transmission tomography. In *Proc. IEEE Conf. Acoust. Speech Sig. Proc.*, volume 4, pages 2291–2294, 1995.

[5] M. R. Segal, P. Bacchetti, and N. P. Jewell. Variances for maximum penalized likelihood estimates obtained via the EM algorithm. *J. Royal Stat. Soc. Ser. B*, 56(2):345–352, 1994.

[6] D. D. Cox and F. O'Sullivan. Asymptotic analysis of penalized likelihood and related estimators. *Ann. Stat.*, 18(4):1676–95, 1990.

[7] F. O'Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1(4):502–27, 1986.

[8] I. B. Kerfoot and Y. Bresler. Theoretical analysis of an information theoretic algorithm for vector field segmentation. *IEEE Tr. Im. Proc.*, 8(6):798–820, June 1999.

[9] D. S. Lalush and B. M. W. Tsui. A fast and stable maximum a posteriori conjugate gradient reconstruction algorithm. *Med. Phys.*, 22(8):1273–84, August 1995.

[10] Z. Liang. Compensation for attenuation, scatter, and detector response in SPECT reconstruction via iterative FBP methods. *Med. Phys.*, 20(4):1097–106, July 1993.

[11] H. M. Hudson and R. S. Larkin. Accelerated image reconstruction using ordered subsets of projection data. *IEEE Tr. Med. Im.*, 13(4):601–9, December 1994.

[12] L. Kaufman. Maximum likelihood, least squares, and penalized least squares for PET. *IEEE Tr. Med. Im.*, 12(2):200–14, June 1993.

[13] H. J. Trussel. Convergence criteria for iterative restoration methods. *IEEE Tr. Acoust. Sp. Sig. Proc.*, 31(1):129–36, February 1983.

[14] H. H. Barrett, D. W. Wilson, and B. M. W. Tsui. Noise properties of the EM algorithm: I. Theory. *Phys. Med. Biol.*, 39(5):833–46, May 1994.

[15] D. W. Wilson, B. M. W. Tsui, and H. H. Barrett. Noise properties of the EM algorithm: II. Monte Carlo simulations. *Phys. Med. Biol.*, 39(5):847–72, May 1994.

[16] D. G. Luenberger. *Optimization by vector space methods*. Wiley, New York, 1969.

[17] C. R. Rao. *Linear statistical inference and its applications*. Wiley, New York, 1973.

[18] R. E. Williamson, R. H. Crowell, and H. F. Trotter. *Calculus of vector functions*. Prentice-Hall, New Jersey, 1972.

[19] D. M. Young. *Iterative solution of large linear systems*. Academic Press, New York, 1971.

[20] R. E. Carson, Y. Yan, B. Chodkowski, T. K. Yap, and M. E. Daube-Witherspoon. Precision and accuracy of regional radioactivity quantitation using the maximum likelihood EM reconstruction algorithm. *IEEE Tr. Med. Im.*, 13(3):526–37, September 1994.

[21] K. Lange and R. Carson. EM reconstruction algorithms for emission and transmission tomography. *J. Comp. Assisted Tomo.*, 8(2):306–16, April 1984.

[22] K. Sauer and C. Bouman. A local update strategy for iterative reconstruction from projections. *IEEE Tr. Sig. Proc.*, 41(2):534–48, February 1993.

[23] J. A. Fessler. Hybrid Poisson/polynomial objective functions for tomographic image reconstruction from transmission scans. *IEEE Tr. Im. Proc.*, 4(10):1439–50, October 1995.

[24] C. Bouman and K. Sauer. Fast numerical methods for emission and transmission tomographic reconstruction. In *Proc. 27th Conf. Info. Sci. Sys., Johns Hopkins*, pages 611–6, 1993.

[25] C. A. Bouman and K. Sauer. A unified approach to statistical tomography using coordinate descent optimization. *IEEE Tr. Im. Proc.*, 5(3):480–92, March 1996.

[26] N. H. Clinthorne, T. S. Pan, P. C. Chiao, W. L. Rogers, and J. A. Stamos. Preconditioning methods for improved convergence rates in iterative reconstructions. *IEEE Tr. Med. Im.*, 12(1):78–83, March 1993.

[27] R. H. Huesman. A new fast algorithm for the evaluation of regions of interest and statistical uncertainty in computed tomography. *Phys. Med. Biol.*, 29(5):543–52, 1984.

[28] J. A. Fessler and A. O. Hero. Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms. *IEEE Tr. Im. Proc.*, 4(10):1417–29, October 1995.

[29] K. Lange. Convergence of EM image reconstruction algorithms with Gibbs smoothing. *IEEE Tr. Med. Im.*, 9(4):439–46, December 1990. Corrections, T-MI, 10:2(288), June 1991.

[30] A. C. Kak and M. Slaney. *Principles of computerized tomographic imaging*. IEEE Press, New York, 1988.

[31] J. A. Fessler. Penalized weighted least-squares image reconstruction for positron emission tomography. *IEEE Tr. Med. Im.*, 13(2):290–300, June 1994.

[32] D. S. Lalush and B. M. W. Tsui. A fast and stable weighted-least squares MAP conjugate-gradient algorithm for SPECT. *J. Nuc. Med. (Abs. Book)*, 34(5):27, May 1993.

[33] J. D. Gorman and A. O. Hero. Lower bounds for parametric estimators with constraints. *IEEE Tr. Info. Theory*, 36(6):1285–1301, November 1990.

[34] T. L. Marzetta. A simple derivation of the constrained multiple parameter Cramer-Rao bound. *IEEE Tr. Sig. Proc.*, 41(6):2247–9, June 1993.

[35] R. E. Carson, Y. Yan, M. E. Daube-Witherspoon, N. Freedman, S. L. Bacharach, and P. Herscovitch. An approximation formula for the variance of PET region-of-interest values. *IEEE Tr. Med. Im.*, 12(2):240–50, June 1993.

[36] P. J. Huber. *Robust statistics*. Wiley, New York, 1981.

[37] J. A. Fessler. Resolution properties of regularized image reconstruction methods. Technical Report 297, Comm. and Sign. Proc. Lab., Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 48109-2122, August 1995.

[38] D. L. Snyder, A. M. Hammoud, and R. L. White. Image recovery from data acquired with a charge-couple-device camera. *J. Opt. Soc. Am. A*, 10(5):1014–23, May 1993.

	Objective	
Term	Likelihood	Quadratic
$\Phi(\theta, Y)$	$\sum_n Y_n \log \bar{Y}_n(\theta) - \bar{Y}_n(\theta)$	$-\frac{1}{2} \sum_n (Y_n - \bar{Y}_n(\theta))^2 / Y_n$
$\frac{\partial}{\partial\theta} \Phi(\theta, Y)$	$\sum_n a_n (Y_n / \bar{Y}_n(\theta) - 1)$	$\sum_n a_n (1 - \bar{Y}_n(\theta) / Y_n)$
$-\frac{\partial^2}{\partial\theta^2} \Phi(\theta, Y)$	$\sum_n a_n^2 Y_n / \bar{Y}_n(\theta)^2$	$\sum_n a_n^2 / Y_n$
$\frac{\partial^3}{\partial\theta^3} \Phi(\theta, Y)$	$\sum_n 2 a_n^3 Y_n / \bar{Y}_n(\theta)^3$	0
$\frac{\partial^2}{\partial\theta \partial Y_n} \Phi(\theta, Y)$	$a_n / \bar{Y}_n(\theta)$	$a_n \bar{Y}_n(\theta) / Y_n^2$
$\frac{\partial^3}{\partial\theta \partial Y_n^2} \Phi(\theta, Y)$	0	$-2 a_n \bar{Y}_n(\theta) / Y_n^3$
$\frac{\partial^3}{\partial\theta^2 \partial Y_n} \Phi(\theta, Y)$	$-a_n^2 / \bar{Y}_n(\theta)^2$	a_n^2 / Y_n^2
$-\frac{\partial^2}{\partial\theta^2} \Phi(\theta, \bar{Y})$	F_θ	F_θ
$\frac{\partial^3}{\partial\theta^3} \Phi(\theta, \bar{Y})$	$2 \sum_n a_n^3 / \bar{Y}_n^2$	0
$\frac{\partial^2}{\partial\theta \partial Y_n} \Phi(\theta, \bar{Y})$	$a_n / \bar{Y}_n$	$a_n / \bar{Y}_n$
$\frac{\partial^3}{\partial\theta \partial Y_n^2} \Phi(\theta, \bar{Y})$	0	$-2 a_n / \bar{Y}_n^2$
$\frac{\partial^3}{\partial\theta^2 \partial Y_n} \Phi(\theta, \bar{Y})$	$-a_n^2 / \bar{Y}_n^2$	$a_n^2 / \bar{Y}_n^2$
$\frac{\partial}{\partial Y_n} h(\bar{Y})$	$a_n / (\bar{Y}_n F_\theta)$	$a_n / (\bar{Y}_n F_\theta)$
$\frac{\partial^2}{\partial Y_n^2} h(\bar{Y})$	$\frac{2}{F_\theta^2} \frac{a_n^2}{\bar{Y}_n^2} \left[\frac{1}{F_\theta} \sum_n a_n^3 / \bar{Y}_n^2 - a_n / \bar{Y}_n \right]$	$\frac{2}{F_\theta} \frac{a_n}{\bar{Y}_n^2} \left[\frac{a_n^2}{\bar{Y}_n F_\theta} - 1 \right]$
$\mathrm{Var}\{\hat{\theta}\} \approx$	$1 / F_\theta$	$1 / F_\theta$
$E\{\hat{\theta}\} - \theta \approx$	0	$\frac{1}{F_\theta^2} \sum_n \frac{a_n^3}{\bar{Y}_n^2} - \frac{1}{F_\theta} \sum_n \frac{a_n}{\bar{Y}_n}$

Table 1: Objective functions and partial derivatives for scalar emission tomography problem with $\bar{Y}_n(\theta) = a_n \theta$.

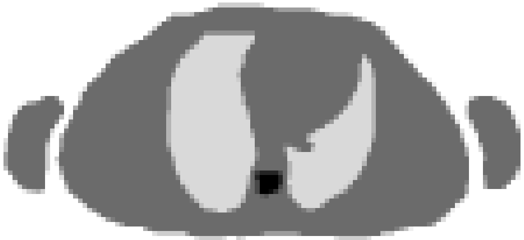

Figure 1: Simulated thorax attenuation map used to evaluate the mean and variance approximations for penalized-likelihood estimators in transmission tomography.

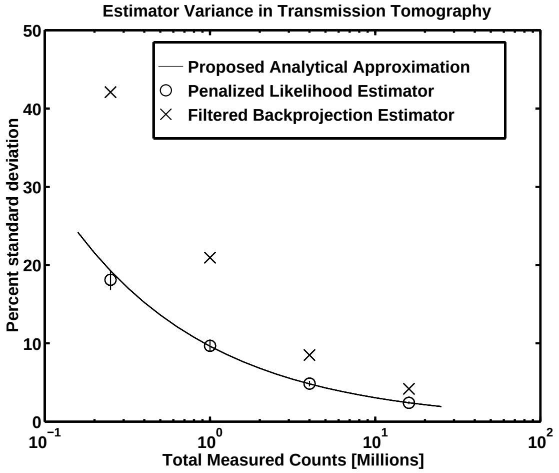

Figure 2: Variance for center pixel of attenuation map as predicted by (26) compared with simulation results from penalized-likelihood estimator (25). Also shown is the variance of conventional FBP.

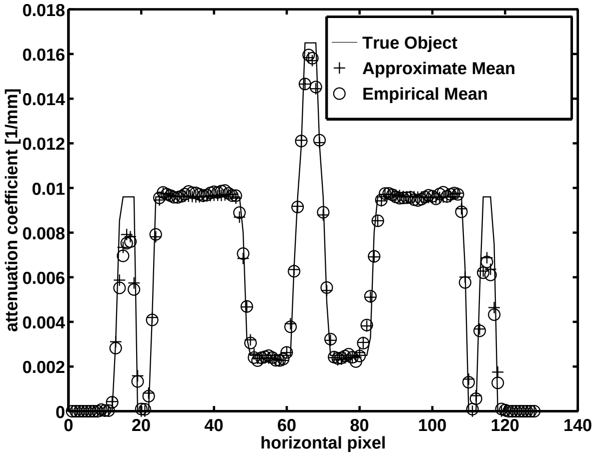

Figure 3: Horizontal cross-section through predicted estimator mean and empirical sample mean. Despite the nonlinearity of the estimator, the prediction agrees closely with the empirical performance.

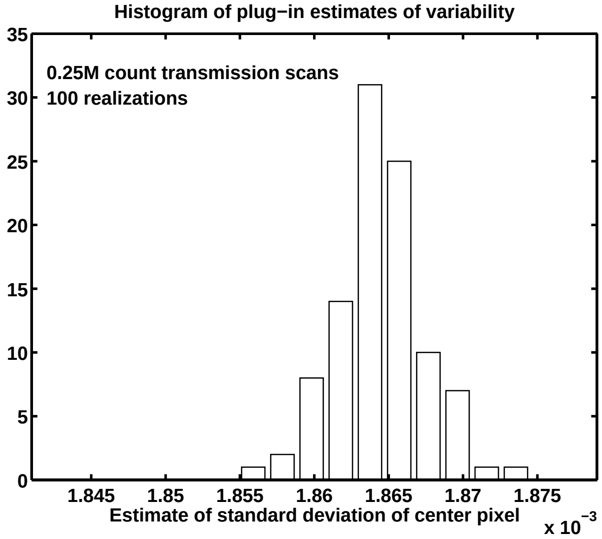

Figure 4: Histogram of 100 post-reconstruction plug-in estimates of variability $\mathrm{Var}\{\theta_j\}$ described by (28), where j corresponds to the center pixel of the attenuation map shown in Fig. 1, for 250K count measurements. The empirical standard deviation from 100 realizations was $1.74 \cdot 10^{-3}\mathrm{mm}^{-1}$.

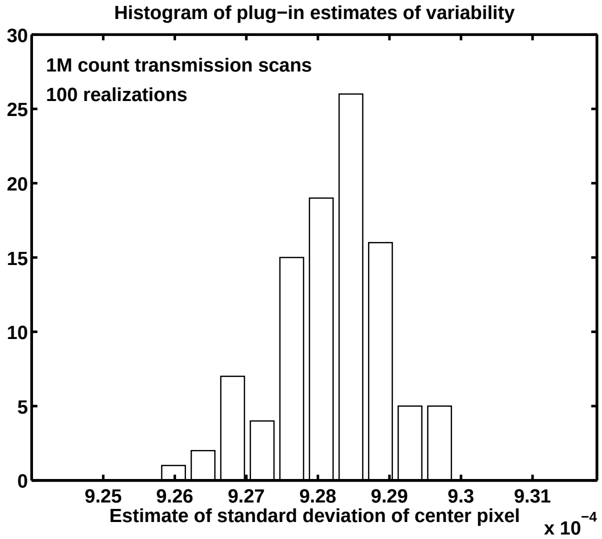

Figure 5: As in previous figure, but for 1M count measurements. The empirical standard deviation from 100 realizations was $9.30 \cdot 10^{-4}\mathrm{mm}^{-1}$.

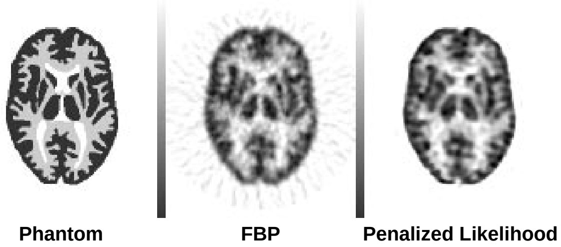

Figure 6: Simulated brain radioisotope emission distribution.

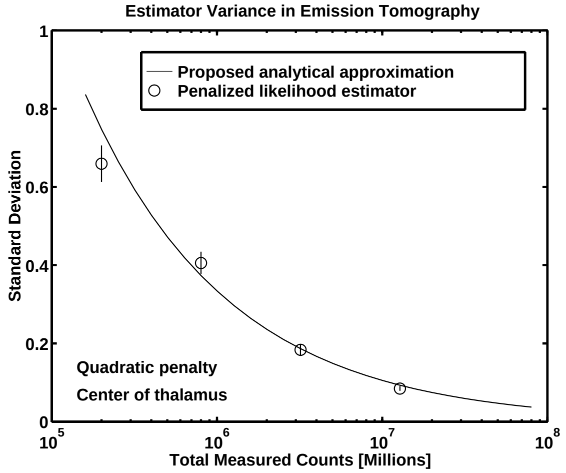

Figure 7: Comparison of predicted variance from (30) with empirical performance of penalized-likelihood emission image reconstruction with quadratic penalty for pixel at center of thalamus.

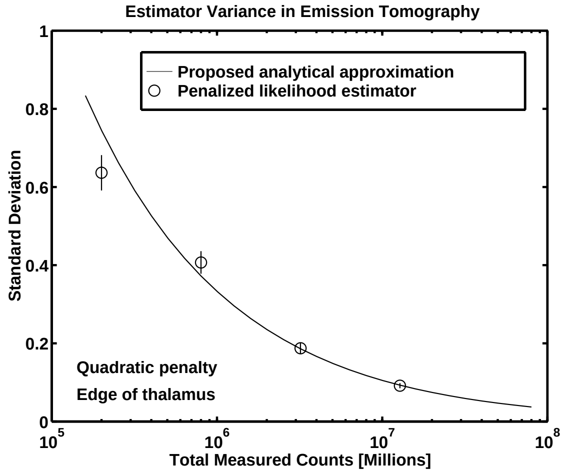

Figure 8: As in Fig. 7 but for pixel at edge of thalamus.

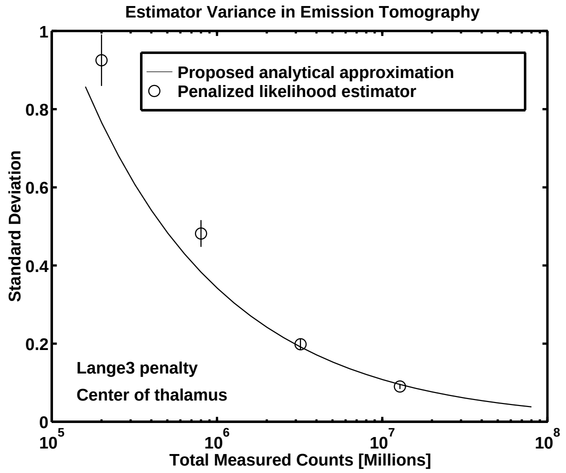

Figure 9: As in Fig. 7 but for nonquadratic penalty (see text).

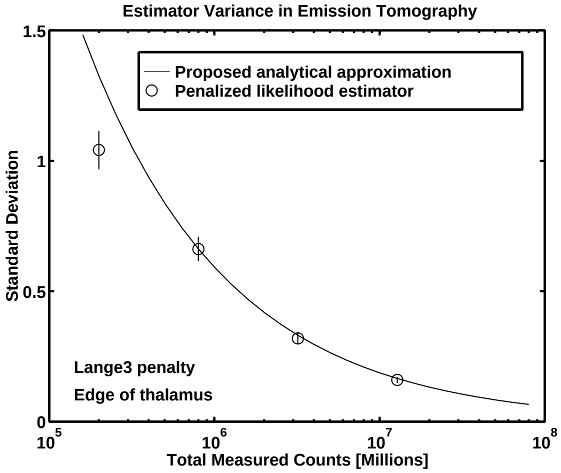

Figure 10: As in Fig. 8 but for nonquadratic penalty (see text).

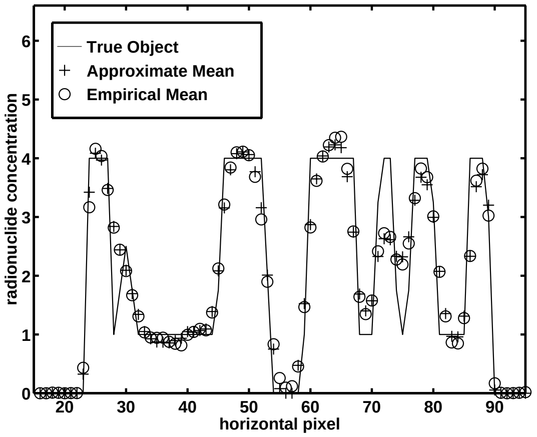

Figure 11: Horizontal profile through emission phantom, 0th-order predicted mean, and empirical mean from penalized-likelihood estimator using nonquadratic penalty for 0.8M count case.

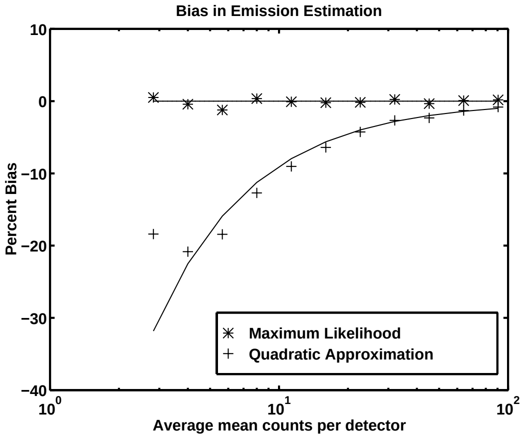

Figure 12: Bias for scalar emission estimation problem for the maximum-likelihood estimator and for the weighted least-squares estimator based on a quadratic approximation to the log-likelihood. Solid lines are the analytical formulas in the last row of Table I; the other points are empirical results.

List of Figures

List of Tables