To appear in Proc. Intl. Conf. Acoust. Speech Signal Proc. (ICASSP) 1995

MOMENTS OF IMPLICITLY DEFINED ESTIMATORS (e.g. ML and MAP): APPLICATIONS TO TRANSMISSION TOMOGRAPHY

Jeffrey A. Fessler

3480 Kresge III, Box 0552
University of Michigan
Ann Arbor, MI 48109-0552
fessler@umich.edu

ABSTRACT

Many estimators in signal processing problems are defined implicitly as the maximum of an objective function, such as maximum likelihood (ML) and maximum a posteriori (MAP) methods. Exact analytical expressions for the mean and variance of such estimators are usually unavailable, so investigators usually resort to numerical simulations. This paper describes approximate analytical expressions for the mean and variance of implicitly defined estimators. The expressions are defined solely in terms of the partial derivatives of whatever objective function one uses for estimation. We demonstrate the utility and accuracy of the approximations in a PET transmission computed tomography application with Poisson statistics. The approximations should be useful in a wide range of estimation problems.

1. INTRODUCTION

Let $\theta = [\theta_1, \ldots, \theta_p]'$ be a unknown parameter that is to be estimated from measurements $Y = [Y_1, \ldots, Y_N]'$, where $'$ denotes vector or matrix transpose. In many areas of signal and image processing, one estimates θ by maximizing some functional of θ and Y:

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

Examples include ML, MAP, penalized maximum likelihood methods, and linear or nonlinear least-squares methods. Except in very simple cases there is usually no explicit analytical expression for $\hat{\theta}$ in terms of Y. In other words, the objective function (1) only *implicitly* defines $\hat{\theta}$ as a function of Y.

The absence of an explicit analytical expression of the form $\hat{\theta} = g(Y)$ 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 maximum 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 [1,2] 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. Expressions for the variance would be particularly useful since having a rough idea of the variance would allow one to determine how many realizations are needed to achieve a desired accuracy in subsequent simulations.

In [3] we used the implicit function theorem, the chain rule, and Taylor's expansion to derive approximate expressions for the mean and variance of implicitly defined estimators of continuous parameters. Here we summarize only the variance expression. We demonstrate the utility and accuracy of the variance approximation for the problem of tomographically reconstructing attenuation images from PET transmission scans with Poisson statistics.

2. THEORY

We restrict our attention to suitably regular objective functions for which one can find the required maximum in (1) by zeroing the partial derivatives of $\Phi(\cdot, Y)$:

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

Thus θ must be a continuous parameter, so this approach is inapplicable to discrete classification prob-

This work was supported in part by NIH Grant CA60711 and DOE Grant DE-FG02-87ER65061.

lems such as image segmentation.

By the implicit function theorem, the relationship (2) implicitly defines a function $\hat{\theta} = h(Y)$ that maps the measurement Y into an estimate of θ. From (2) the function $h(Y)$ must satisfy

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

Perhaps slightly abusing notation, we rewrite (3) as:

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

where we 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 arguments 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 exactly compute its mean or variance since the estimator $h(Y)$ is usually nonlinear. Although exact 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

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

Taking the expectation of both sides yields the following well-known approximation for the mean:

$$E\{\hat{\theta}\} \approx h(\bar{Y}) + \frac{1}{2} \sum_n \sum_m \frac{\partial^2}{\partial Y_n \partial Y_m} h(\bar{Y}) \mathrm{Cov}\{Y_n, Y_m\}, \qquad (5)$$

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, which we assume is known. Taking the covariance of both sides of the first-order Taylor expansion of $h(Y)$ yields the following approximation for the covariance:

$$\mathrm{Cov}\{\hat{\theta}\} \approx \nabla h(\bar{Y}) \; \mathrm{Cov}\{Y\} \; [\nabla h(\bar{Y})]', \qquad (6)$$

where $\nabla = [\frac{\partial}{\partial Y_1} \; \ldots \; \frac{\partial}{\partial Y_N}]$ denotes the (row) gradient operator.

Since $h(Y)$ is unknown, (5) and (6) are not immediately applicable. However, (5) and (6) depend on $h(Y)$ only through its partial derivatives[1]. From the calculus of vector functions [4, p. 302], one can determine the partial derivatives of an implicit function by applying the chain rule with respect to Y_n to (4):

$$\begin{aligned} 0 = \sum_k \frac{\partial^2}{\partial \theta_j \partial \theta_k} \Phi(h(Y), Y) \; \frac{\partial}{\partial Y_n} h_k(Y) \\ + \frac{\partial^2}{\partial \theta_j \partial Y_n} \Phi(h(Y), Y), j = 1, \ldots, p. \end{aligned} \qquad (7)$$

This expression has N sets of p equations in p unknowns which can be written in matrix form:

$$0 = \nabla^{20} \Phi(h(Y), Y) \; \nabla h(Y) + \nabla^{11} \Phi(h(Y), Y), \qquad (8)$$

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 $\frac{\partial^2}{\partial \theta_j \partial Y_n}$. Assuming the symmetric matrix $-\nabla^{20} \Phi(h(Y), Y)$ is also positive definite, we can solve for ∇g by rearranging:

$$\nabla h(Y) = [-\nabla^{20} \Phi(h(Y), Y)]^{-1} \; \nabla^{11} \Phi(h(Y), Y). \qquad (9)$$

Let $\tilde{\theta} = h(\bar{Y})$ and combine (9) with (6) to obtain the following approximation to the covariance of $\hat{\theta}$:

$$\begin{aligned} \mathrm{Cov}\{\hat{\theta}\} \;&\approx\; [-\nabla^{20} \Phi(\tilde{\theta}, \bar{Y})]^{-1} \; \nabla^{11} \Phi(\tilde{\theta}, \bar{Y}) \; \mathrm{Cov}\{Y\} \\ &\cdot [\nabla^{11} \Phi(\tilde{\theta}, \bar{Y})]' \; [-\nabla^{20} \Phi(\tilde{\theta}, \bar{Y})]^{-1}. \end{aligned} \qquad (10)$$

This final expression is an approximation to the estimator covariance that *depends only on the partial derivatives of the objective function* $\Phi(\theta, Y)$, and not on the implicit function $h(Y)$.

When p is large, the full covariance matrix becomes inconvenient to store, and often one is interested primarily in the variance of certain parameters in a region of interest. Let e be the jth unit vector of length p, and define $u = [-\nabla^{20} \Phi(\tilde{\theta}, \bar{Y})]^{-1} e$. Note that one does not need to perform a $p \times p$ matrix inversion to compute u; one simply needs to solve the equation $[-\nabla^{20} \Phi(\tilde{\theta}, \bar{Y})] u = e$. From (10) it follows that

$$\mathrm{Var}\{\hat{\theta}_j\} \approx u' \; \nabla^{11} \Phi(\tilde{\theta}, \bar{Y}) \; \mathrm{Cov}\{Y\} \; [\nabla^{11} \Phi(\tilde{\theta}, \bar{Y})]' u. \qquad (11)$$

In many imaging problems, the covariance of Y is diagonal and the partial derivatives of Φ are sparse (see below), so the actual computation is comparable to two image reconstructions/restorations.

By differentiating (8) again, we similarly obtain approximate expressions for the estimator mean (or bias), as shown in [3].

[1]Except that (5) also depends on $h(\bar{Y})$. If we define $\tilde{\theta} = \arg\max_\theta \Phi(\theta, \bar{Y})$, then from (4), we see that $h(\bar{Y}) = \tilde{\theta}$, so one can easily find $h(\bar{Y})$ by applying the estimation algorithm to the noise free data $\bar{Y}$.

3. APPLICATION TO TOMOGRAPHY

To illustrate the accuracy of the approximation for estimator covariance given by (10), 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 [5–7]. 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:

$$\begin{aligned} \bar{Y}_n(\theta) &= T p_n(\theta) \\ p_n(\theta) &= b_n e^{-\sum_j a_{nj}\theta_j} + r_n, \end{aligned} \tag{12}$$

where the a_{nj} factors denote the intersection length of the nth ray passing though the jth pixel, and T denotes the scan duration. The nonnegative factors b_n, r_n, and a_{nj} are assumed known. The log-likelihood is:

$$L(\theta, Y) = \sum_n Y_n \log \bar{Y}_n(\theta) - \bar{Y}_n(\theta), \tag{13}$$

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

$$\Phi(\theta, Y) = \frac{1}{T} L(\theta, Y) - \beta V(\theta), \tag{14}$$

where V is a roughness penalty of the form

$$V(\theta) = \sum_j \frac{1}{2} \sum_k w_{jk} \phi(\theta_j - \theta_k), \tag{15}$$

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

Due to the nonlinearity of (12) and the non-quadratic likelihood function (13) for Poisson statistics, the estimate $\hat{\theta}$ formed by maximizing (14) 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 variance approximation (10). We focus on the case where $r_n = 0$.

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 (11), which requires the following partial derivatives:

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

where $\mathbf{R}(\theta) = \nabla^2 V(\theta)$ is the matrix of second partials of $V(\theta)$.

Since the measurements have independent Poisson distributions, it follows that $\mathrm{Cov}\{Y\} = \mathrm{diag}\left\{\bar{Y}_n(\theta^{\mathrm{true}})\right\}$. Substituting into (10) and simplifying yields the following approximation to the estimator covariance:

$$\mathrm{Cov}\{\hat{\theta}\} \approx \frac{1}{T}(\mathbf{F}(\tilde{\theta}) + \beta \mathbf{R}(\tilde{\theta}))^{-1} \mathbf{F}(\theta^{\mathrm{true}}) (\mathbf{F}(\tilde{\theta}) + \beta \mathbf{R}(\tilde{\theta}))^{-1}, \tag{16}$$

where $\mathbf{F}(\theta) = \mathbf{A}' \mathrm{diag}\{p_n(\theta)\} \mathbf{A}$ is $1/T$ times the Fisher information for estimating θ from Y.

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

- Compute $\tilde{\theta} = \arg\max_\theta \Phi(\theta, \bar{Y})$ by applying to noise free data $\bar{Y}$ a maximization algorithm such as coordinate ascent [7,8].
- Forward project $\tilde{\theta}$ to compute $p_n(\tilde{\theta}) = \sum_j a_{nj}\tilde{\theta}_j$. Likewise for $p_n(\theta^{\mathrm{true}})$.
- Pick a pixel j of interest and solve the equation $(\mathbf{A}'\mathrm{diag}\left\{p_n(\tilde{\theta})\right\}\mathbf{A} + \beta\mathbf{R}(\tilde{\theta}))u = e$ for u using a fast iterative method such as preconditioned conjugate gradients or Gauss-Siedel [9].
- Compute $\frac{1}{T} u' \mathbf{A}' \mathrm{diag}\{p_n(\theta^{\mathrm{true}})\} \mathbf{A} u$ by first forward projecting u to compute $q = \mathbf{A}u$, and then summing:

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

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.

To assess the accuracy of the above approximation, we performed numerical simulations using a synthetic human thorax cross-section attenuation map as

θ^{true}, with linear attenuation coefficients 0.0165/mm, 0.0096/mm, and 0.0025/mm, 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 correspond to 6mm wide strip integrals on 3mm center-to-center spacing. (This is an approximation to the ideal line integral that accounts for finite detector width.) The b_n factors were generated using pseudo-random log-normal variates with a standard deviation of 0.3 to account for detector efficiency variations. Four studies were performed, with the scale factor T set so that $\sum_n \bar{Y}_n(\theta^{\text{true}})$ was 0.25, 1, 4, and 16 million counts. The r_n factors were set to 0 for simplicity. For each study, 100 realizations of pseudo-random Poisson transmission measurements were generated according to (12) and then reconstructed using the penalized likelihood estimator described by (14) using a coordinate-ascent algorithm [7, 8]. The coordinate-ascent algorithm enforced the nonnegativity constraint $\hat{\theta} \geq 0$. For simplicity, we used the function $\phi(x) = x^2/2$ for the penalty in (15). 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 (16). Fig. 1 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 (16) 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.

4. FOR MORE INFORMATION

Preprints including [1, 3, 7] are available using WWW from the following URL.

`http://www.engin.umich.edu/~fessler/`

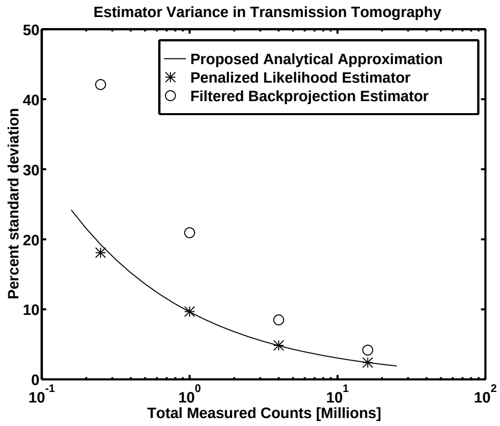

Figure 1: Variance for center pixel as predicted by (16) compared with simulation results from a penalized likelihood estimation algorithm (14). Also shown is the variance of conventional FBP.

5. REFERENCES

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

[2] M Usman, A O Hero, J A Fessler, and W L Rogers. Bias-variance tradeoffs analysis using uniform CR bound for a SPECT system. In *Conf. Rec. of the IEEE Nuc. Sci. Symp. Med. Im. Conf.*, pp. 1463–1467, 1993.

[3] J A Fessler. Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): applications to tomography, 1994. Submitted to IEEE Trans. Image Proc.

[4] R E Williamson, R H Crowell, and H F Trotter. *Calculus of Vector Functions.* Prentice Hall, New Jersey, 1972.

[5] K Lange and R Carson. EM reconstruction algorithms for emission and transmission tomography. *J. Comp. Assisted Tomo.*, 8(2):306–316, April 1984.

[6] K Sauer and C Bouman. A local update strategy for iterative reconstruction from projections. *IEEE Trans. Sig. Proc.*, 41(2):534–548, Feb. 1993.

[7] J A Fessler. Hybrid Poisson/polynomial objective functions for tomographic image reconstruction from transmission scans. *IEEE Trans. Im. Proc.*, 1994. To appear.

[8] C Bouman and K Sauer. Fast numerical methods for emission and transmission tomographic reconstruction. In *Proc. 27th Conf. Info. Sci. Sys., Johns Hopkins*, pp. 611–616, 1993.

[9] D M Young. *Iterative Solution of Large Linear Systems.* Academic Press, New York, 1971.