Chapter 23

Mean and Variance Analysis

ch,mav

Contents

s,mav,intro

23.1 Introduction (s,mav,intro)

Most of the reconstruction methods described in this book are estimators defined *implicitly* as the minimizer of some cost function, such as penalized-likelihood methods. For such estimators, exact analytical expressions for statistical properties such as the mean and covariance are usually unavailable[1]. In contrast, one can easily analyze the statistics of *linear* reconstruction methods such as the FBP method [2] in tomography or the conjugate phase method in MRI. Thus, investigators often resort to numerical simulations to examine statistical properties of nonlinear estimators. Although empirical studies are important, analytical expressions, even if approximate, can be convenient for comparing estimators, for designing imaging systems, and for developing intuition.

In §1.9 we analyzed the covariance of certain simple statistical methods for image restoration. This chapter describes approximations for the mean and covariance of more general implicitly defined estimators of *unconstrained* continuous parameters. We derive the approximations using the *implicit function theorem* [3, p. 266], the *Taylor expansion*, and the *chain rule*. The expressions are defined solely in terms of the partial derivatives of whatever cost function one uses for estimation. Simulations demonstrating that the approximations work well in two tomographic imaging applications are given in [4]. The approximations are useful in a wide range of estimation problems. This chapter is based largely on [4], but also includes a more accurate approximation described in [5].

23.2 Background

Let $\boldsymbol{x} = (x_1, \ldots, x_{n_\mathrm{p}}) \in \mathbb{R}^{n_\mathrm{p}}$ denote a unknown real parameter vector that is to be estimated from a real measurement vector $\boldsymbol{y} = (y_1, \ldots, y_{n_\mathrm{d}}) \in \mathbb{R}^{n_\mathrm{d}}$. For many image reconstruction problems, one specifies an estimator $\hat{\boldsymbol{x}}$ to be the minimizer of some cost function:

e,mav,cost

$$\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}(\boldsymbol{y}) = \arg\min_{\boldsymbol{x}} \Psi(\boldsymbol{x}, \boldsymbol{y}) \,. \tag{23.2.1}$$

Examples of such methods include *maximum-likelihood* (*ML*) estimation, *maximum* a posteriori (*MAP*) 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 expresses $\hat{\boldsymbol{x}}$ explicitly in terms of $\boldsymbol{y}$. In other words, the cost function (23.2.1) defines $\hat{\boldsymbol{x}}$ only *implicitly* as a function of $\boldsymbol{y}$. Statisticians refer to (23.2.1) as an *M-estimate* [6].

The absence of an explicit analytical expression for $\hat{\boldsymbol{x}}(\boldsymbol{y})$ makes it difficult to study the mean and covariance of the estimator $\hat{\boldsymbol{x}}$, 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 n_p is large). Similar considerations apply if one wishes to compare estimator performance against the uniform *Cramér-Rao* bound for biased estimators (see §23.18) to examine the bias-variance trade-off of the estimator [7], [8]. Therefore, it is useful to have approximate expressions for the mean and covariance of implicitly defined estimators, particularly if those approximations require less computation than multiple numerical simulations [4], [9].

For unbiased maximum-likelihood estimation, the *Cramér-Rao* bound can serve as an approximation to the estimator variance. But *bias* is unavoidable for regularized methods, so the unbiased *Cramér-Rao* bound is inapplicable. In the statistics literature, approximate covariances for penalized-likelihood estimates have been computed for specific iterative algorithms [10], but most analyses of penalized-likelihood methods have focused on the asymptotic properties of mean squared error *e.g.*, [11], [12]. For practical signal-to-noise ratios, bias and variance may have unequal importance in imaging problems, in contrast to their equal weighting in the mean squared error performance measure.

In this chapter we apply the implicit function theorem, the Taylor expansion, and the chain rule to (23.2.1) to derive approximate expressions for the mean and covariance of implicitly defined estimators $\hat{\boldsymbol{x}}$. 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.

The expressions are similar to the asymptotic moments given by Serfling [6] for scalar M-estimates. Our focus here is on presenting a simple derivation of useful approximations for multi-parameter imaging problems, rather than on asymptotic analysis.

Because of the partial derivatives used in the derivation, our approximations are restricted to problems where $\boldsymbol{x}$ is a continuous parameter vector. Thus the approach is inapplicable to discrete classification problems such as image

[1] Even in the cases described in [1] where the exact distribution of the estimator $\hat{\boldsymbol{x}}$ has been found, integrating that expression to find the mean or covariance remains an *open problem*.

segmentation. (Mean and variance are poor performance measures for segmentation problems anyway; analyses of classification errors are more appropriate [13].) Furthermore, strictly speaking we must also exclude problems where inequality constraints are imposed on $\hat{\boldsymbol{x}}$, because when the minimization in (23.2.1) is subject to inequality constraints, one must replace (23.3.2) below with appropriate *Karush-Kuhn-Tucker* (*KKT*) conditions. Our focus is on imaging problems, where often the only inequality constraint is nonnegativity of $\hat{\boldsymbol{x}}$. This constraint is often particularly important in *unregularized* estimation methods. However, for cost functions that include a regularization term, our experience is that nonnegativity constraints are active relatively infrequently. So the variances of the unconstrained and constrained estimators are approximately equal for most pixels (*cf.* [14]). Qi and Leahy proposed generalizations that use truncated gaussian distributions to account for nonnegativity [15], [16].

The derivations assume the estimate is computed by "completely" minimizing the cost function, *i.e.*, the approximations are not applicable to unregularized methods for which one uses a "stopping rule" to terminate the iterations long before the minimizer is reached. In particular, our results are inapplicable to unregularized methods such as the iterative filter-backproject method [17] and the ordered subsets expectation maximization (*OSEM*) algorithm [18]. Except in simple linear cases [19], it is generally difficult to analyze the performance of methods based on stopping rules, although Barrett et al. [20], [21] have analyzed the per-iteration behavior of the maximum-likelihood expectation maximization (*MLEM*) algorithm for emission tomography, and this analysis has been generalized for other iterations [22]–[24]. The approximations we derive are somewhat easier to use because they are independent of number of iterations (provided sufficient iterations are used to minimize the cost function).

23.3 Covariance of implicit estimators (s,mav,cov)

s,mav,cov

Let $\Psi(\boldsymbol{x}, \boldsymbol{y})$ denote a cost function that depends on unknown parameters $\boldsymbol{x}$ and noisy measurements $\boldsymbol{y}$. Define an estimator $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}(\boldsymbol{y})$ as the (unconstrained) minimizer of this cost function, as in (23.2.1). We assume $\Psi(\cdot, \boldsymbol{y})$ has a unique global minimizer $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}(\boldsymbol{y}) \in \mathbb{R}^{n_\mathrm{p}}$ for any measurement $\boldsymbol{y}$, so that $\hat{\boldsymbol{x}}(\boldsymbol{y})$ is a well defined function. We also assume that Ψ is suitably regular that the partial derivatives used below exist. This section describes approximations for the covariance of the estimator $\hat{\boldsymbol{x}}$.

23.3.1 Covariance approximation 1

Define the $n_\mathrm{p} \times 1$ column gradient of the cost function following the notation in (22.3.2):

e,mav,kgrad

$$\boldsymbol{\Gamma}(\boldsymbol{x}, \boldsymbol{y}) \triangleq \nabla^{[1,0]} \Psi(\boldsymbol{x}, \boldsymbol{y}) \,. \tag{23.3.1}$$

Then a necessary condition for the minimizer $\hat{\boldsymbol{x}}$ is that it is a zero of the gradient:

e,mav,d10=0

$$\boldsymbol{\Gamma}(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) = \boldsymbol{0}. \tag{23.3.2}$$

This requires that Ψ be suitably regular, and it is this step that restricts our approximations to continuous parameters and that precludes inequality constraints and stopping rules. For generalizations to the constrained case, see [25].

Let $\check{\boldsymbol{x}}$ denote some non-random nominal value for the parameter vector, such as $\boldsymbol{x}_\mathrm{true}$, and make a *first-order Taylor series* expansion (27.8.3) of $\boldsymbol{\Gamma}$ around $\check{\boldsymbol{x}}$:

e,mav,kgrad,approx

$$\boldsymbol{\Gamma}(\hat{\boldsymbol{x}}, \boldsymbol{y}) \approx \boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \boldsymbol{y}) + \nabla^{[1,0]} \boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \boldsymbol{y}) \, (\hat{\boldsymbol{x}} - \check{\boldsymbol{x}}) \,, \tag{23.3.3}$$

where the $n_\mathrm{p} \times n_\mathrm{p}$ matrix $\nabla^{[1,0]} \boldsymbol{\Gamma}$ is the Hessian of Ψ:

$$\nabla^{[1,0]} \boldsymbol{\Gamma} = \nabla^{[2,0]} \Psi \,.$$

In statistics, (23.3.3) is known as the *Delta method* [26, Ch. 3]. Using (23.3.2), we equate the approximation (23.3.3) to zero, yielding:

$$\boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \boldsymbol{y}) \approx -\nabla^{[1,0]} \boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \boldsymbol{y}) \, (\hat{\boldsymbol{x}} - \check{\boldsymbol{x}}) = -\nabla^{[2,0]} \Psi(\check{\boldsymbol{x}}, \boldsymbol{y}) \, (\hat{\boldsymbol{x}} - \check{\boldsymbol{x}}) \,.$$

Rearranging yields the following linearized approximations for the estimator:

$$\begin{aligned} \hat{\boldsymbol{x}} &\approx \check{\boldsymbol{x}} - \left[\nabla^{[2,0]} \Psi(\check{\boldsymbol{x}}, \boldsymbol{y}) \right]^{-1} \boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \boldsymbol{y}) \\ &\approx \check{\boldsymbol{x}} - \boldsymbol{H}^{-1} \boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \boldsymbol{y}), \end{aligned} \tag{23.3.4}$$

assuming that the Hessian of Ψ is invertible[2], where $\boldsymbol{H}$ denotes some (non-random) approximation to the Hessian of Ψ. This $\boldsymbol{H}$ may depend on $\boldsymbol{x}_\mathrm{true}$ or $\check{\boldsymbol{x}}$, but not on the random measurement $\boldsymbol{y}$. Taking the covariance of both sides of (23.3.4) yields the following covariance approximation

e,mav,cov,1

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \approx \boldsymbol{H}^{-1} \, \mathsf{Cov}\{\boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \boldsymbol{y})\} \, \boldsymbol{H}^{-1}. \tag{23.3.5}$$

Similar expressions are found in [6], [27, p. 133], [26, p. 52], for the asymptotic covariances of *M-estimators*. The scalar case appeared in [28]. The practical utility of this approximation hinges on how easily one can compute or approximate $\mathsf{Cov}\{\boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \boldsymbol{y})\}$, and on the choice for $\check{\boldsymbol{x}}$. These considerations are discussed further below.

[2] For example, if Ψ has a positive definite Hessian, then $\boldsymbol{H}$ will be invertible and Ψ will be strictly convex.

23.3.2 Choosing the linearization point $\check{\boldsymbol{x}}$

The accuracy of the covariance approximation (23.3.5) will depend on how one chooses the linearization point $\check{\boldsymbol{x}}$ and the Hessian approximation $\boldsymbol{H}$. To have the correct asymptotic properties [27, p. 133] [26, p. 52], the best choice for $\check{\boldsymbol{x}}$ is

e,mav,xch,ex

$$\check{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \mathsf{E}[\Psi(\boldsymbol{x}, \boldsymbol{y})] = \arg\min_{\boldsymbol{x}} \int \Psi(\boldsymbol{x}, \boldsymbol{y}) \, \mathsf{p}(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}, \tag{23.3.6}$$

where $\mathsf{p}(\boldsymbol{y})$ denotes the distribution of the data $\boldsymbol{y}$. In general, this $\check{\boldsymbol{x}}$ could be difficult to compute. Fortunately, some cost functions are *effectively affine* in $\boldsymbol{y}$, meaning they are affine to within a possibly $\boldsymbol{y}$-dependent constant that is independent of $\boldsymbol{x}$ and hence does not affect the minimizer[3]. Examples include the negative log-likelihood for the gaussian and Poisson statistical models. In such cases, $\check{\boldsymbol{x}}$ simplifies as follows:

e,mav,xch,yb

$$\check{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \Psi(\boldsymbol{x}, \mathsf{E}[\boldsymbol{y}]) = \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}), \tag{23.3.7}$$

where $\bar{\boldsymbol{y}} = \mathsf{E}[\boldsymbol{y}]$ denotes the mean measurements, *i.e.*, noiseless data. I used this latter choice for $\check{\boldsymbol{x}}$ in [4] as an approximation even for cases where Ψ is nonlinear in $\boldsymbol{y}$.

23.3.3 Choosing the Hessian approximation $\boldsymbol{H}$

Similarly, to ensure that (23.3.5) has the appropriate asymptotic properties [27, p. 133], we should define $\boldsymbol{H}$ as follows:

e,mav,cov,H,ex

$$\boldsymbol{H} = \mathsf{E}\left[\nabla^{[1,0]} \boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \boldsymbol{y})\right] = \mathsf{E}\left[\nabla^{[2,0]} \, \Psi(\check{\boldsymbol{x}}, \boldsymbol{y})\right]. \tag{23.3.8}$$

With this choice, the estimates $\hat{\boldsymbol{x}}$ have a normal distribution asymptotically with mean $\check{\boldsymbol{x}}$ and covariance given by (23.3.5) [26, p. 52]. When the cost function Ψ is chosen as the negative log-likelihood, this Hessian $\boldsymbol{H}$ is simply the *Fisher information*. However, it can be difficult to compute $\boldsymbol{H}$ in (23.3.8) in some situations. Fortunately, for cost functions that are effectively affine in $\boldsymbol{y}$, the Hessian simplifies as follows:

e,mav,cov,H,hess

$$\boldsymbol{H} = \nabla^{[1,0]} \boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \mathsf{E}[\boldsymbol{y}]) = \nabla^{[2,0]} \, \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}). \tag{23.3.9}$$

For other cases, we can consider the preceding $\boldsymbol{H}$ to be an approximation that should be accurate if the cost function is *approximately* effectively affine in $\boldsymbol{y}$. Substituting into (23.3.5) yields the following covariance approximation

e,mav,cov,1H

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \approx \left[\nabla^{[2,0]} \, \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})\right]^{-1} \mathsf{Cov}\{\boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \boldsymbol{y})\} \left[\nabla^{[2,0]} \, \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})\right]^{-1}. \tag{23.3.10}$$

This form may still be inconvenient due to the middle term.

23.3.4 Covariance approximation 2

As a further approximation, we could linearize $\boldsymbol{\Gamma}$ around the mean measurements:

e,mav,cov,kgrad,lin

$$\boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \boldsymbol{y}) \approx \boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}) + \nabla^{[0,1]} \boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})(\boldsymbol{y} - \bar{\boldsymbol{y}}), \tag{23.3.11}$$

where (see (22.3.3)):

$$\nabla^{[0,1]} \boldsymbol{\Gamma}(\boldsymbol{x}, \boldsymbol{y}) = \nabla^{[1,1]} \, \Psi(\boldsymbol{x}, \boldsymbol{y}).$$

The approximation (23.3.11) is exact for cost functions that are effectively affine in $\boldsymbol{y}$. This linearization suggests the second approximation:

$$\mathsf{Cov}\{\boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \boldsymbol{y})\} \approx [\nabla^{[1,1]} \, \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})] \, \mathsf{Cov}\{\boldsymbol{y}\} [\nabla^{[1,1]} \, \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})]',$$

which, substituted into (23.3.5), yields the following covariance approximation:

e,mav,cov,2

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \approx \left[\nabla^{[2,0]} \, \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})\right]^{-1} \left[\nabla^{[1,1]} \, \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})\right] \mathsf{Cov}\{\boldsymbol{y}\} \left[\nabla^{[1,1]} \, \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})\right]' \left[\nabla^{[2,0]} \, \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})\right]^{-1}. \tag{23.3.12}$$

Clearly (23.3.12) has an additional level of approximation beyond that in (23.3.5), namely, it assumes a degree of linearity in the effect of the measurements $\boldsymbol{y}$ on $\boldsymbol{\Gamma}$ that may be unrealistic. So (23.3.5) is the preferable approximation when its use is feasible.

[3] In other words, the gradient $\nabla^{[1,0]} \, \Psi(\boldsymbol{x}, \boldsymbol{y})$ is affine in $\boldsymbol{y}$.

s,mav,cov,lin

23.3.5 Covariance approximation 3 (s,mav,cov,lin)

In (23.3.3), we used a first-order Taylor expansion of the gradient function $\boldsymbol{\Gamma}$ to derive the covariance approximation (23.3.5). An alternative approach is to use a first-order Taylor expansion of the estimator itself:

e,mav,cov,lin,xh

$$\hat{\boldsymbol{x}}(\boldsymbol{y}) \approx \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) + \nabla \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) \, (\boldsymbol{y} - \bar{\boldsymbol{y}}) \,, \tag{23.3.13}$$

where $\bar{\boldsymbol{y}}$ was defined below (23.3.9). Taking the covariance[4] of both sides yields the following well-known approximation [29, p. 426]:

e,mav,cov,taylor

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} = \mathsf{Cov}\{\hat{\boldsymbol{x}}(\boldsymbol{y})\} \approx \nabla \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) \; \mathsf{Cov}\{\boldsymbol{y}\} \; \nabla \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) \,. \tag{23.3.14}$$

If we knew $\hat{\boldsymbol{x}}(\cdot)$ then we could apply (23.3.14) directly to approximate the covariance of $\hat{\boldsymbol{x}}(\boldsymbol{y})$. But because $\hat{\boldsymbol{x}}$ is unknown, (23.3.14) is not immediately useful. However, the dependence on $\hat{\boldsymbol{x}}$ in (23.3.14) is *only through its partial derivatives* at the point $\bar{\boldsymbol{y}}$. From the calculus of vector functions [30, p. 302], one can determine the partial derivatives of an implicitly defined function by applying the chain rule, as shown in (22.3.6):

e,mav,grad,xh

$$\nabla \hat{\boldsymbol{x}}(\boldsymbol{y}) = \left[\nabla^{[2,0]} \, \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y})\right]^{-1} \left[-\nabla^{[1,1]} \, \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y})\right] . \tag{23.3.15}$$

Substituting into (23.3.14) yields the same approximation as (23.3.12) [4].

To summarize, (23.3.5) and (23.3.12) are the main results here: approximate expressions for the estimator covariance that depend only on the partial derivatives of the cost function Ψ, and do not require an expression for the implicit function $\hat{\boldsymbol{x}}(\boldsymbol{y})$. These approximations do depend on $\check{\boldsymbol{x}}$, which one usually computes using (23.3.7) by applying the estimation algorithm to the noise free data $\bar{\boldsymbol{y}}$.

s,mav,pl

23.3.6 Penalized-likelihood estimators (s,mav,pl)

By analogy with §22.4, focus now on *penalized-likelihood* estimators of the form (22.4.1):

$$\Psi(\boldsymbol{x}, \boldsymbol{y}) = \text{Ł}(\boldsymbol{x}, \boldsymbol{y}) + \mathsf{R}(\boldsymbol{x}) \,.$$

Then the covariance approximation (23.3.12) specializes to be

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \approx \left[\nabla^{[2,0]} \, \text{Ł}(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}) + \mathbf{R}\right]^{-1} \left(\nabla^{[1,1]} \, \text{Ł}(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})\right) \mathsf{Cov}\{\boldsymbol{y}\} \left(\nabla^{[1,1]} \, \text{Ł}(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})\right)' \left[\nabla^{[2,0]} \, \text{Ł}(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}) + \mathbf{R}\right]^{-1} ,$$

where the penalty Hessian is $\mathbf{R} = \nabla^2 \, \mathsf{R}(\check{\boldsymbol{x}})$.

23.3.6.1 Independent measurements

To further simplify, follow §22.4.1 and consider the usual case where the measurements $\{y_i\}$ are statistically independent and the negative log-likelihood has the form (22.4.6):

$$\text{Ł}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n_\mathrm{d}} g_i(\bar{y}_i(\boldsymbol{x}), y_i) \,.$$

Using the assumption (22.4.8) that leads to the log-likelihood gradient expressions (22.4.9) and substituting into the covariance expression above yields

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \approx \left[\boldsymbol{B}' \boldsymbol{D}_2(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \boldsymbol{B} + \mathbf{R}\right]^{-1} \boldsymbol{B}' \boldsymbol{D}_1(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \, \mathsf{Cov}\{\boldsymbol{y}\} \, \boldsymbol{D}_1(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \boldsymbol{B} \left[\boldsymbol{B}' \boldsymbol{D}_2(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \boldsymbol{B} + \mathbf{R}\right]^{-1} ,$$

where $\boldsymbol{B} = \nabla \bar{\boldsymbol{y}}(\check{\boldsymbol{x}})$ and $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$ were defined in terms of the derivatives of the g_i functions in (22.4.10).

Usually $\boldsymbol{D}_1 \, \mathsf{Cov}\{\boldsymbol{y}\} \, \boldsymbol{D}_1 \approx \boldsymbol{D}_2$. and the preceding covariance approximation simplifies to the form

e,mav,pl,ind

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \approx \left[\mathbf{F} + \mathbf{R}\right]^{-1} \mathbf{F} \left[\mathbf{F} + \mathbf{R}\right]^{-1} , \tag{23.3.16}$$

where $\mathbf{F} = \boldsymbol{B}' \boldsymbol{D}_2 \boldsymbol{B}$ is the Fisher information for estimating $\boldsymbol{x}$ from $\boldsymbol{y}$, evaluated at $\check{\boldsymbol{x}}$.

23.3.6.2 Image reconstruction problems

Following Problem 22.4.2, for image reconstruction problems usually (22.4.13) holds:

$$g_i(\bar{y}_i(\boldsymbol{x}), y_i) = \mathsf{h}_i([\boldsymbol{A}\boldsymbol{x}]_i \,, y_i)$$

for some functions $\{\mathsf{h}_i\}$. In this case $\boldsymbol{B} = \boldsymbol{A}$ and $\boldsymbol{D}_2$ is defined in terms of h_i in (22.4.14). The covariance approximation (23.3.16) still holds.

[4] All expectations and covariances are taken with respect to the probability density of the random measurement $\boldsymbol{y}$. Typically one assumes this density is of the form $\mathsf{p}(\boldsymbol{y}; \boldsymbol{x}_\mathrm{true})$, where $\boldsymbol{x}_\mathrm{true}$ is the unknown parameter to be estimated using (23.2.1). 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 §23.11).

s,mav,cov,roi,intro

23.4 Computing covariances within a region (s,mav,cov,roi,intro)

When n_p is large, storing the full $n_\mathrm{p} \times n_\mathrm{p}$ covariance matrix is inconvenient, and often one is interested primarily in the variance or covariance of certain parameters in a *region of interest* (*ROI*). Specifically, often we want to evaluate

$$\mathsf{Cov}\{\hat{x}_j, \hat{x}_l\} = \boldsymbol{e}_j' \, \mathsf{Cov}\{\hat{\boldsymbol{x}}\} \, \boldsymbol{e}_l$$

for j, l within some ROI.

s,mav,cov,roi,mat

23.4.1 Matrix approach (s,mav,cov,roi,mat)

Let $\boldsymbol{e}_j$ be the jth unit vector of length n_p, and define

e,mav,cov,roi,mat,inv

$$\boldsymbol{u}_j = \left[\nabla^{[2,0]} \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})\right]^{-1} \boldsymbol{e}_j. \tag{23.4.1}$$

Note that one does not need to perform a $n_\mathrm{p} \times n_\mathrm{p}$ matrix inversion to compute $\boldsymbol{u}_j$; one "simply" solves the equation $\left[\nabla^{[2,0]} \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})\right] \boldsymbol{u}_j = \boldsymbol{e}_j$. This can be done directly when n_p is small, or via fast iterative methods such as Gauss-Siedel or conjugate gradients when n_p is large [31]. From (23.3.12) it follows that

$$\begin{aligned} \mathsf{Cov}\{\hat{x}_j, \hat{x}_l\} &= \boldsymbol{e}_j' \, \mathsf{Cov}\{\hat{\boldsymbol{x}}\} \, \boldsymbol{e}_l \\ &\approx \boldsymbol{u}_j' \left[\nabla^{[1,1]} \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})\right] \mathsf{Cov}\{\boldsymbol{y}\} \left[\nabla^{[1,1]} \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}})\right]' \boldsymbol{u}_l, \end{aligned} \tag{23.4.2}$$

for $l, j = 1, \ldots, n_\mathrm{p}$. One can compute any portion of the covariance matrix of $\hat{\boldsymbol{x}}$ by using (23.4.2) repeatedly for appropriate j and l values. In general, computing $\mathsf{Var}\{\hat{x}_j\} = \mathsf{Cov}\{\hat{x}_j, \hat{x}_j\}$ for all j using this formula requires $O(n_\mathrm{p}^2 + n_\mathrm{d} n_\mathrm{p} + n_\mathrm{d}^2)$ operations. In many problems, such as the tomographic examples in §23.10 and §23.12, the covariance of $\boldsymbol{y}$ is diagonal and the partial derivatives have a sparse structure, so the actual computation can be much less. Nevertheless, evaluating (23.4.1) requires about about as much work as doing iterative reconstruction, just to predict the variance of a single voxel or ROI. Therefore, we describe simpler approximations next.

s,mav,cov,roi,fft

23.4.2 FFT-based approach (s,mav,cov,roi,fft)

As described in §22.7, for problems that are locally shift invariant near pixel j_0, we can make "locally circulant" approximations for $j \approx j_0$:

$$\begin{aligned} \boldsymbol{T}\boldsymbol{H}\boldsymbol{e}_j &\approx \boldsymbol{Q}^{-1} \, \mathsf{diag}\{\mathsf{H}_{j_0,k}\} \, \boldsymbol{Q}\boldsymbol{T}\boldsymbol{e}_j \\ \boldsymbol{T}\boldsymbol{J}\boldsymbol{e}_j &\approx \boldsymbol{Q}^{-1} \, \mathsf{diag}\{\mathsf{J}_{j_0,k}\} \, \boldsymbol{Q}\boldsymbol{T}\boldsymbol{e}_j, \end{aligned}$$

where

$$\begin{aligned} \boldsymbol{H} &= \nabla^{[2,0]} \Psi \\ \boldsymbol{J} &= \left[\nabla^{[1,1]} \Psi\right] \mathsf{Cov}\{\boldsymbol{y}\} \left[\nabla^{[1,1]} \Psi\right]' \\ \mathsf{diag}\{\mathsf{H}_{j_0,k}\} &= \mathsf{diag}\{\boldsymbol{Q}\boldsymbol{T}\boldsymbol{H}\boldsymbol{e}_{j_0}\} \\ \mathsf{diag}\{\mathsf{J}_{j_0,k}\} &= \mathsf{diag}\{\boldsymbol{Q}\boldsymbol{T}\boldsymbol{J}\boldsymbol{e}_{j_0}\}, \end{aligned}$$

and where $\boldsymbol{T}$ is the $N_\mathrm{p} \times n_\mathrm{p}$ matrix defined in §22.7, where in 2D we have $N_\mathrm{p} = NM$.

Substituting these approximations into (23.3.12) and simplifying yields

e,mav,cov,roi,fft,cov

$$\mathsf{Cov}\{\hat{x}_j, \hat{x}_l\} \approx \boldsymbol{e}_j' \boldsymbol{Q}^{-1} \, \mathsf{diag}\left\{\mathsf{J}_{j_0,k} / (\mathsf{H}_{j_0,k})^2\right\} \boldsymbol{Q}\boldsymbol{e}_l = r_{j_0}[j-l], \tag{23.4.3}$$

for $j, l \approx j_0$, where the auto-correlation function $r_{j_0}[n]$ is the inverse DFT of the *local power spectral density* function $\mathsf{J}_{j_0,k} / (\mathsf{H}_{j_0,k})^2$. In particular, the approximate variance is

$$\mathsf{Var}\{\hat{x}_j\} = \boldsymbol{e}_j' \, \mathsf{Cov}\{\hat{x}_j, \hat{x}_l\} \, \boldsymbol{e}_j \approx \boldsymbol{e}_j' \boldsymbol{Q}^{-1} \, \mathsf{diag}\left\{\mathsf{J}_{j_0,k} / (\mathsf{H}_{j_0,k})^2\right\} \boldsymbol{Q}\boldsymbol{e}_j = \frac{1}{N_\mathrm{p}} \sum_k \mathsf{J}_{j_0,k} / (\mathsf{H}_{j_0,k})^2.$$

We used the property that the DFT of an impulse is unity, *i.e.*, $\boldsymbol{Q}\boldsymbol{e}_j = \mathbf{1}$.

For typical penalized-likelihood cost functions, usually $\boldsymbol{H} = \mathsf{F} + \mathsf{R}$ and $\boldsymbol{J} = \mathsf{F}$, so the variance expression becomes

e,mav,cov,roi,fft,var

$$\mathsf{Var}\{\hat{x}_j\} = \boldsymbol{e}_j' \boldsymbol{Q}^{-1} \, \mathsf{diag}\left\{\frac{\mathsf{F}_{j_0,k}}{(\mathsf{F}_{j_0,k} + \mathsf{R}_{j_0,k})^2}\right\} \boldsymbol{Q}\boldsymbol{e}_j = \frac{1}{N_\mathrm{p}} \sum_k \frac{\mathsf{F}_{j_0,k}}{(\mathsf{F}_{j_0,k} + \mathsf{R}_{j_0,k})^2}, \tag{23.4.4}$$

for $j \approx j_0$, where $\mathsf{F}_{j_0,k}$ and $\mathsf{R}_{j_0,k}$ are defined by

$$\begin{aligned} \mathsf{diag}\{\mathsf{F}_{j_0,k}\} &= \boldsymbol{Q}\boldsymbol{T}\mathbf{F}\boldsymbol{e}_{j_0} \\ \mathsf{diag}\{\mathsf{R}_{j_0,k}\} &= \boldsymbol{Q}\mathbf{R}\mathbf{F}\boldsymbol{e}_{j_0}. \end{aligned} \tag{23.4.5}$$

Usually $\mathbf{F} = \boldsymbol{A}'\boldsymbol{W}\boldsymbol{A}$ for some diagonal matrix $\boldsymbol{W}$. After computing a forward and backprojection of a single pixel to compute $\mathbf{F}\boldsymbol{e}_{j_0}$ and taking the FFT to form $\mathsf{F}_{j_0,k}$, and a similar operation to find $\mathsf{R}_{j_0,k}$, one can compute the predicted variance or any given voxel for a range of values of the regularization parameter β using just a few FFT operations.

The variance expression (23.4.4) generalizes the analysis for image restoration (1.9.7) in several ways. First it does not require $\boldsymbol{A}$ to be circulant, but rather only that $\mathbf{F}$ is locally shift invariant. Secondly it allows for nonuniform measurement noise statistics. Finally, it depends on the choice of the reference pixel index j_0, so it can characterize non-stationary estimator statistics.

IRT *See* `qpwls_psf.m` *for examples.*

s,mav,roi,plug

23.4.3 Plug-in approximation for noisy data (s,mav,roi,plug)

The approximation (23.3.12) for the estimator covariance depends on both $\check{\boldsymbol{x}}$ and $\mathsf{Cov}\{\boldsymbol{y}\}$, so as written its primary use is in computer simulations where $\check{\boldsymbol{x}}$ and $\mathsf{Cov}\{\boldsymbol{y}\}$ are known. Sometimes one would like to obtain an approximate estimate of estimator variability from a single noisy measurement (such as real data), for which $\boldsymbol{x}_{\mathrm{true}}$ is unknown, and $\mathsf{Cov}\{\boldsymbol{y}\}$ may also be unknown. In some problems this can be done using a "*plug-in*" estimate in which we substitute the estimate $\hat{\boldsymbol{x}}$ in for $\check{\boldsymbol{x}}$ in expressions like (23.3.12). The effectiveness of this approach is application dependent, but was shown to be useful for transmission tomography in [4]. It is also useful in emission tomography provided one is careful with some details [32].

s,mav,var

23.5 Fast variance map predictions (s,mav,var)

Using the FFT-based variance approximation (23.4.4) is practical when one is interested in the noise of only a few pixels. But if one wants to form a *variance map* for the entire image, then still faster methods are desired.

We focus on the usual case where $\mathbf{F} = \boldsymbol{A}'\boldsymbol{W}\boldsymbol{A}$ where $\boldsymbol{W} = \mathsf{diag}\{w_i\}$ is a diagonal matrix whose entries might depend on $\check{\boldsymbol{x}}$ and $\bar{y}_i$.

23.5.1 Tabulation approach based on certainty factors

We first simplify by making the approximation (see (22.9.3)) [33]:

$$\mathbf{F} = \boldsymbol{A}'\boldsymbol{W}\boldsymbol{A} \approx \boldsymbol{D}\boldsymbol{A}'\boldsymbol{A}\boldsymbol{D}$$

where $\boldsymbol{D} = \mathsf{diag}\{\kappa_j\}$ and (see (22.9.2))

$$\kappa_j \triangleq \sqrt{\frac{\sum_{i=1}^{n_\mathrm{d}} a_{ij}^2 w_i}{\sum_{i=1}^{n_\mathrm{d}} a_{ij}^2}} \approx \sqrt{\frac{\sum_{i=1}^{n_\mathrm{d}} a_{ij} w_i}{\sum_{i=1}^{n_\mathrm{d}} a_{ij}}}.$$

Assume that *local* to the jth voxel the penalty Hessian is approximately

$$\mathbf{R} \approx \beta_j \boldsymbol{C}'\boldsymbol{C}$$

for a differencing matrix $\boldsymbol{C}$. Then

$$\mathsf{Var}\{\hat{x}_j\} \approx \frac{1}{\kappa_j^2}\boldsymbol{e}_j' \left[\boldsymbol{A}'\boldsymbol{A} + \frac{\beta_j}{\kappa_j^2}\mathbf{R}\right]^{-1} \boldsymbol{A}'\boldsymbol{A} \left[\boldsymbol{A}'\boldsymbol{A} + \frac{\beta_j}{\kappa_j^2}\mathbf{R}\right]^{-1} \boldsymbol{e}_j = \frac{1}{\kappa_j^2}\sigma^2\left(\frac{\beta_j}{\kappa_j^2}\right), \tag{23.5.1}$$

e,mav,var,cert

where

$$\sigma^2(\beta) \triangleq \boldsymbol{e}_j' \left[\boldsymbol{A}'\boldsymbol{A} + \beta\boldsymbol{C}'\boldsymbol{C}\right]^{-1} \boldsymbol{A}'\boldsymbol{A} \left[\boldsymbol{A}'\boldsymbol{A} + \beta\boldsymbol{C}'\boldsymbol{C}\right]^{-1} \boldsymbol{e}_j.$$

We can compute the table $\sigma^2(\cdot)$ using circulant approximations, as described in §23.4.2.

If we use the modified regularizer of [33] (see §22.10), then $\beta_j = \beta_0\kappa_j^2$ and the variance approximation simplifies to

$$\mathsf{Var}\{\hat{x}_j\} \approx \frac{1}{\kappa_j^2}\sigma^2(\beta_0), \tag{23.5.2}$$

e,mav,var,kapj

i.e., the noise standard deviation is inversely proportional κ_j. Therefore, using a modified regularizer to strive for uniform spatial resolution can lead to nonuniform noise. There appears to be a trade-off between uniform noise and uniform spatial resolution in systems with nonstationary measurement noise statistics like PET and X-ray CT.

23.5.2 Analytical approximations

In some applications we can find fast approximations by replacing summations with integrals. See [34], [35] for examples in X-ray CT.

s,mav,joint

23.6 Joint estimation (s,mav,joint)

In a variety of applications we perform joint estimation of an image $\boldsymbol{z}$ and some other system-model parameters $\boldsymbol{\alpha}$ using a cost function of the following form

$$\Psi(\boldsymbol{z};\boldsymbol{y}) = \Psi((\boldsymbol{x},\boldsymbol{\alpha});\boldsymbol{y}) = \frac{1}{2}\|\boldsymbol{y}-\boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x}\|^2_{\boldsymbol{W}^{1/2}} + \mathsf{R}_1(\boldsymbol{x}) + \mathsf{R}_2(\boldsymbol{\alpha}),$$

where $\boldsymbol{z} = (\boldsymbol{x},\boldsymbol{\alpha})$. Assuming that $\bar{\boldsymbol{y}}(\boldsymbol{z}) = \bar{\boldsymbol{y}}(\boldsymbol{x},\boldsymbol{\alpha}) = \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x}$ and $\mathsf{Cov}\{\boldsymbol{y}\} = \boldsymbol{W}^{-1}$, and further assuming the noise is gaussian, the Fisher information matrix for this estimation problem is

$$\mathbf{F} = \mathsf{E}[\nabla \mathsf{L} \nabla \mathsf{L}] = (\nabla \bar{\boldsymbol{y}})' \boldsymbol{W} (\nabla \bar{\boldsymbol{y}}),$$

where if $\boldsymbol{\alpha} \in \mathbb{R}^K$ then the gradient matrix is $n_\mathrm{d} \times (n_\mathrm{p} + K)$ as follows:

$$\nabla \bar{\boldsymbol{y}} = \left[\begin{array}{cc} \boldsymbol{A} & \nabla_{\boldsymbol{\alpha}} \bar{\boldsymbol{y}} \end{array} \right],$$

where $\nabla_{\boldsymbol{\alpha}} \bar{\boldsymbol{y}} = \nabla_{\boldsymbol{\alpha}} \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x}$ is the $n_\mathrm{d} \times K$ matrix with elements

$$\frac{\partial}{\partial \alpha_k} [\boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x}]_i = \sum_{j=1}^{n_\mathrm{p}} \frac{\partial}{\partial \alpha_k} a_{ij}(\boldsymbol{\alpha}) x_j, \quad i = 1, \ldots, n_\mathrm{d}, \quad k = 1, \ldots, K.$$

This expression does not appear to simplify further in general, because $\frac{\partial}{\partial \alpha_k} a_{ij}(\boldsymbol{\alpha})$ depends on three indices. Therefore the $(n_\mathrm{p} + K) \times (n_\mathrm{p} + K)$ Fisher information matrix has the following block form:

$$\mathbf{F} = \left[\begin{array}{cc} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{12}' & \mathbf{F}_{22} \end{array} \right] = \left[\begin{array}{cc} \boldsymbol{A}'\boldsymbol{W}\boldsymbol{A} & \boldsymbol{A}'\boldsymbol{W}(\nabla_{\boldsymbol{\alpha}} \bar{\boldsymbol{y}}) \\ \mathbf{F}_{12}' & (\nabla_{\boldsymbol{\alpha}} \bar{\boldsymbol{y}})' \boldsymbol{W} (\nabla_{\boldsymbol{\alpha}} \bar{\boldsymbol{y}}) \end{array} \right].$$

The matrix $\mathbf{F}_{11}$ is the Fisher information for estimating $\boldsymbol{x}$ assuming $\boldsymbol{\alpha}$ is known, whereas $\mathbf{F}_{22}$ is the Fisher information for estimating $\boldsymbol{\alpha}$ assuming $\boldsymbol{x}$ is known.

In the presence of regularization, the estimators can be biased so the inverse of the Fisher information is not an accurate approximation to the estimator covariance. We can apply the covariance approximation (23.3.12) to examine the properties of estimates of the combined parameter vector $(\boldsymbol{x},\boldsymbol{\alpha})$. First note that

$$\nabla^{[1,0]} \Psi = \nabla_{\boldsymbol{z}} \Psi = \left[\begin{array}{c} -\boldsymbol{A}'\boldsymbol{W}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}) + \nabla \mathsf{R}_1(\boldsymbol{x}) \\ -(\nabla_{\boldsymbol{\alpha}} \bar{\boldsymbol{y}})' \boldsymbol{W}(\boldsymbol{y}-\boldsymbol{A}\boldsymbol{x}) + \nabla \mathsf{R}_2(\boldsymbol{\alpha}) \end{array} \right],$$

so it follows that

$$-\nabla^{[1,1]} \Psi = \nabla_{\boldsymbol{y}} \nabla_{\boldsymbol{z}} \Psi = \left[\begin{array}{c} \boldsymbol{A}'\boldsymbol{W} \\ (\nabla_{\boldsymbol{\alpha}} \bar{\boldsymbol{y}})' \boldsymbol{W} \end{array} \right],$$

which is a $(n_\mathrm{p} + K) \times n_\mathrm{d}$ matrix. Because $\mathsf{Cov}\{\boldsymbol{y}\} = \boldsymbol{W}^{-1}$, the middle three terms in (23.3.12) simplify to

$$\left(-\nabla^{[1,1]} \Psi\right) \mathsf{Cov}\{\boldsymbol{y}\} \left(-\nabla^{[1,1]} \Psi\right)' = \mathbf{F}.$$

In addition

$$\nabla^{[2,0]} \Psi = \nabla^2_{\boldsymbol{z}} \Psi = \left[\begin{array}{cc} \boldsymbol{H}_{11} & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{12}' & \boldsymbol{H}_{22} \end{array} \right] = \left[\begin{array}{cc} \mathbf{F}_{11} + \mathbf{R}_1 & \boldsymbol{H}_{12} \\ \boldsymbol{H}_{12}' & \tilde{\mathbf{F}}_{22} + \mathbf{R}_2 \end{array} \right],$$

where the elements of $\tilde{\mathbf{F}}_{22}$ are

$$\begin{aligned} [\tilde{\mathbf{F}}_{22}]_{kl} &= \frac{\partial^2}{\partial \alpha_k \partial \alpha_l} \frac{1}{2} \langle \boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x},\ \boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x} \rangle_{\boldsymbol{W}} \\ &= \left\langle \frac{\partial}{\partial \alpha_k} \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x},\ \frac{\partial}{\partial \alpha_l} \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x} \right\rangle_{\boldsymbol{W}} - \left\langle \frac{\partial^2}{\partial \alpha_k \partial \alpha_l} \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x},\ \boldsymbol{y} - \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x} \right\rangle_{\boldsymbol{W}}. \end{aligned}$$

When we use the covariance approximation (23.3.12), we evaluate $\mathbf{F}_{22}$ at $(\check{\boldsymbol{x}},\check{\boldsymbol{\alpha}})$ and $\bar{\boldsymbol{y}}$. Often $\bar{\boldsymbol{y}} \approx \boldsymbol{A}(\check{\boldsymbol{\alpha}})\check{\boldsymbol{x}}$ in which case we can disregard the second term and use the simpler approximation

$$[\tilde{\mathbf{F}}_{22}]_{kl} \approx \left\langle \frac{\partial}{\partial \alpha_k} \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x},\ \frac{\partial}{\partial \alpha_l} \boldsymbol{A}(\boldsymbol{\alpha})\boldsymbol{x} \right\rangle_{\boldsymbol{W}} = [\mathbf{F}_{22}]_{kl}.$$

The $n_\mathrm{p} \times K$ matrix $\boldsymbol{H}_{12}$ is defined by

$$\boldsymbol{H}_{12} = \nabla_{\boldsymbol{\alpha}} \left(\boldsymbol{A}'(\boldsymbol{\alpha}) \boldsymbol{W} \left(\boldsymbol{A}(\boldsymbol{\alpha}) \boldsymbol{x} - \boldsymbol{y} \right) \right),$$

which has elements

$$\begin{aligned} [\boldsymbol{H}_{12}]_{jk} &= \frac{\partial}{\partial \alpha_k} \left[\boldsymbol{A}'(\boldsymbol{\alpha}) \boldsymbol{W} \left(\boldsymbol{A}(\boldsymbol{\alpha}) \boldsymbol{x} - \boldsymbol{y} \right) \right]_j \\ &= \sum_{i=1}^{n_\mathrm{d}} a_{ij} \left[\boldsymbol{W} \left(\frac{\partial}{\partial \alpha_k} \boldsymbol{A}(\boldsymbol{\alpha}) \boldsymbol{x} \right) \right]_i + \sum_{i=1}^{n_\mathrm{d}} \left(\frac{\partial}{\partial \alpha_k} a_{ij}(\boldsymbol{\alpha}) \right) \left[\boldsymbol{W} \left(\bar{\boldsymbol{y}} - \boldsymbol{y} \right) \right]_i . \end{aligned}$$

Again, often we can disregard the second term yielding the simpler approximation $\boldsymbol{H}_{12} \approx \mathsf{F}_{12}$. Combining all of these approximations yields:

$$\mathsf{Cov}\left\{ \begin{bmatrix} \hat{\boldsymbol{x}} \\ \hat{\boldsymbol{\alpha}} \end{bmatrix} \right\} \approx \begin{bmatrix} \mathsf{F}_{11} + \mathsf{R}_1 & \mathsf{F}_{12} \\ \mathsf{F}_{12}' & \mathsf{F}_{22} + \mathsf{R}_2 \end{bmatrix}^{-1} \begin{bmatrix} \mathsf{F}_{11} & \mathsf{F}_{12} \\ \mathsf{F}_{12}' & \mathsf{F}_{22} \end{bmatrix} \begin{bmatrix} \mathsf{F}_{11} + \mathsf{R}_1 & \mathsf{F}_{12} \\ \mathsf{F}_{12}' & \mathsf{F}_{22} + \mathsf{R}_2 \end{bmatrix}^{-1}.$$

For example use of these covariance approximation, see [36, Ch. 4].

23.7 Iteration-dependent covariance (s,mav,iter)

s,mav,iter

Many of the algorithms in this *book* use iterations of the form:

$$\boldsymbol{x}^{(n+1)} = \arg\min_{\boldsymbol{x}} \phi(\boldsymbol{x}, \boldsymbol{x}^{(n)}, \boldsymbol{y}) = M(\boldsymbol{x}^{(n)}, \boldsymbol{y}),$$

where $M : \mathbb{R}^{n_\mathrm{p}} \times \mathbb{R}^{n_\mathrm{d}} \to \mathbb{R}^{n_\mathrm{p}}$. By linearizing the mapping M, one can analyze the noise covariance *as a function of iteration* [37]. First define the "noise free" iteration sequence by

$$\bar{\boldsymbol{x}}^{(n+1)} \triangleq M(\bar{\boldsymbol{x}}^{(n)}, \bar{\boldsymbol{y}})$$

where $\bar{\boldsymbol{x}}^{(0)}$ is a non-random initial estimate such as a uniform image, and $\bar{\boldsymbol{y}}$ is noiseless data. Now linearize M about $\bar{\boldsymbol{x}}^{(n)}$ and $\bar{\boldsymbol{y}}$:

$$\boldsymbol{x}^{(n+1)} = M(\boldsymbol{x}^{(n)}, \boldsymbol{y}) \approx M(\bar{\boldsymbol{x}}^{(n)}, \bar{\boldsymbol{y}}) + (\nabla^{10} M)(\bar{\boldsymbol{x}}^{(n)}, \bar{\boldsymbol{y}})(\boldsymbol{x}^{(n)} - \bar{\boldsymbol{x}}^{(n)}) + (\nabla^{01} M)(\bar{\boldsymbol{x}}^{(n)}, \bar{\boldsymbol{y}})(\boldsymbol{y} - \bar{\boldsymbol{y}}).$$

Let $\boldsymbol{\delta}^{(n)} \triangleq \boldsymbol{x}^{(n)} - \bar{\boldsymbol{x}}^{(n)}$ denote the randomness due to noise at the nth iteration. Then

$$\boldsymbol{\delta}^{(n+1)} \approx (\nabla^{10} M)(\bar{\boldsymbol{x}}^{(n)}, \bar{\boldsymbol{y}}) \boldsymbol{\delta}^{(n)} + (\nabla^{01} M)(\bar{\boldsymbol{x}}^{(n)}, \bar{\boldsymbol{y}})(\boldsymbol{y} - \bar{\boldsymbol{y}}).$$

From this one can derive an expression for the covariance of $\boldsymbol{x}^{(n)}$ at each iteration:

e,mav,cov,iter

$$\mathsf{Cov}\{\boldsymbol{x}^{(n)}\} \approx \boldsymbol{U}_n \, \mathsf{Cov}\{\boldsymbol{y}\} \, \boldsymbol{U}_n', \tag{23.7.1}$$

where $\boldsymbol{U}_n$ is a $n_\mathrm{p} \times n_\mathrm{p}$ matrix that satisfies the following recursive expression:

$$\boldsymbol{U}_{n+1} = (\nabla^{10} M)(\bar{\boldsymbol{x}}^{(n)}, \bar{\boldsymbol{y}}) \boldsymbol{U}_n + (\nabla^{01} M)(\bar{\boldsymbol{x}}^{(n)}, \bar{\boldsymbol{y}}).$$

One can show that the iteration-dependent covariance expression (23.7.1) converges to the "fixed-point" expression (23.3.14) under suitable conditions on ϕ. See Problem 23.2.

23.8 Mean approximations (s,mav,mean)

s,mav,mean

To approximate the mean of $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}(\boldsymbol{y})$, one has several choices, described next.

23.8.1 First-order mean approximation

The simplest approach is to take the expectation of the first-order Taylor expansion (23.3.13), yielding the approximation:

e,mav,mean,1

$$\mathsf{E}[\hat{\boldsymbol{x}}] = \mathsf{E}[\hat{\boldsymbol{x}}(\boldsymbol{y})] \approx \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) . \tag{23.8.1}$$

Interestingly, a 0th-order Taylor expansion yields the same result. This approximation is simply the value produced by applying the estimator (23.2.1) to *noise-free data*. This approach requires modest computation, and often works surprisingly well for penalized-likelihood cost functions. It has been used extensively by investigators in emission tomography [20], [21], [38]. Apparently, the principal source of bias in penalized-likelihood estimators is the regularizing penalty that one includes in Ψ, so (23.8.1) allows one to examine the effects of the penalty separately from the effects of noise. However, the approximation (23.8.1) is certainly not always adequate, as an example in [4] illustrates. Therefore, we next use a second-order Taylor expansion to derive a mean approximation that is more accurate, but has the disadvantage of greater computation.

23.8.2 Second-order mean approximation

A second-order Taylor expansion of the estimator $\hat{\boldsymbol{x}}(\boldsymbol{y})$ around $\bar{\boldsymbol{y}}$ yields:

$$\begin{aligned}\hat{\boldsymbol{x}}(\boldsymbol{y}) \quad &\approx \quad \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) + \sum_{i=1}^{n_\mathrm{d}} \frac{\partial}{\partial y_i} \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) \, (y_i - \bar{y}_i) \\ &\quad + \frac{1}{2} \sum_{i=1}^{n_\mathrm{d}} \sum_{l=1}^{n_\mathrm{d}} \frac{\partial^2}{\partial y_i \, \partial y_l} \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) \, (y_i - \bar{y}_i) \, (y_l - \bar{y}_l) \,. \end{aligned} \tag{23.8.2}$$

Taking the expectation of both sides yields the following well-known approximation for the mean of $\hat{\boldsymbol{x}}(\boldsymbol{y})$:

e,mav,mean,2

$$\mathsf{E}[\hat{\boldsymbol{x}}] \approx \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) + \frac{1}{2} \sum_{i=1}^{n_\mathrm{d}} \sum_{l=1}^{n_\mathrm{d}} \frac{\partial^2}{\partial y_i \, \partial y_l} \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) \, \mathsf{Cov}\{y_i, y_l\}, \tag{23.8.3}$$

where $\mathsf{Cov}\{y_i, y_l\} = \mathsf{E}[(y_i - \bar{y}_i)(y_l - \bar{y}_l)]$ is the (i,l)th element of the covariance matrix of $\boldsymbol{y}$. The approximation (23.8.3) requires the second partial derivatives of $\hat{\boldsymbol{x}}(\boldsymbol{y})$. To obtain those partial derivatives, we first rewrite the equality (22.3.5) as follows:

e,mav,chain1

$$0 = \sum_{k=1}^{n_\mathrm{p}} \frac{\partial^2}{\partial x_j \, \partial x_k} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) \, \frac{\partial}{\partial y_i} \hat{x}_k(\boldsymbol{y}) + \frac{\partial^2}{\partial x_j \, \partial y_i} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) \,. \tag{23.8.4}$$

Next we use the chain rule to differentiate (23.8.4) with respect to y_m, obtaining:

$$\begin{aligned} 0 = &\sum_{k=1}^{n_\mathrm{p}} \left[\sum_{l=1}^{n_\mathrm{p}} \frac{\partial^3}{\partial x_j \partial x_k \partial x_l} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) \, \frac{\partial}{\partial y_m} \hat{x}_l(\boldsymbol{y}) + \frac{\partial^3}{\partial x_j \partial x_k \partial y_m} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) \right] \frac{\partial}{\partial y_i} \hat{x}_k(\boldsymbol{y}) \\ &+ \sum_{k=1}^{n_\mathrm{p}} \frac{\partial^2}{\partial x_j \, \partial x_k} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) \, \frac{\partial^2}{\partial y_i \, \partial y_m} \hat{x}_k(\boldsymbol{y}) + \sum_{k=1}^{n_\mathrm{p}} \frac{\partial^3}{\partial x_j \partial x_k \partial y_i} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) \, \frac{\partial}{\partial y_m} \hat{x}_k(\boldsymbol{y}) \\ &+ \frac{\partial^3}{\partial x_j \partial y_i \partial y_m} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}), \end{aligned} \tag{23.8.5}$$

e,mav,mean,chain2

for $j = 1, \ldots, n_\mathrm{p}$, $i = 1, \ldots, n_\mathrm{d}$, $m = 1, \ldots, n_\mathrm{d}$. One can substitute $\check{\boldsymbol{x}} = \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}})$ and $\boldsymbol{y} = \bar{\boldsymbol{y}}$ in the above expression to obtain n_d^2 sets of n_p equations in the n_p unknowns $\left\{ \frac{\partial^2}{\partial y_i \, \partial y_m} \hat{x}_k(\bar{\boldsymbol{y}}) \right\}_{k=1}^{p}$. Solving each of those systems of equations and then substituting back into (23.8.3) yields an approximation to $\mathsf{E}[\hat{\boldsymbol{x}}]$ that is independent of the unknown implicit function $\hat{\boldsymbol{x}}(\boldsymbol{y})$. If n_p and n_d are large in a given problem, then one must weigh the relative computational expense of solving the above equations versus performing numerical simulations. The trade-off will depend on the structure of the cost function Ψ. Note that (23.8.5) depends on the first partials $\frac{\partial}{\partial y_i} \hat{x}_k(\boldsymbol{y})$, so one must first apply (23.3.15) to compute those partials.

Unlike expression (23.8.4), which we were able to write in the matrix form (23.3.15), there does not appear to be a simple form for rewriting (23.8.5), except by introducing tensor products (which really do not offer much simplification). However, the equations in (23.8.5) do simplify for some special cases for Ψ, described next.

23.8.2.1 Independent measurements

If the measurements $\{y_i\}_{i=1}^{n_\mathrm{d}}$ are statistically independent, then (23.8.3) simplifies to

e,mav,mean,ind

$$\mathsf{E}[\hat{\boldsymbol{x}}(\boldsymbol{y})] \approx \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) + \frac{1}{2} \sum_{i=1}^{n_\mathrm{d}} \frac{\partial^2}{\partial y_i^2} \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) \, \mathsf{Var}\{y_i\} \,. \tag{23.8.6}$$

This expression depends only on the diagonal elements of the covariance of $\boldsymbol{y}$ and on the diagonal of the matrix of second partial derivatives of $\hat{\boldsymbol{x}}(\boldsymbol{y})$. Therefore one needs only the cases where $m = i$ in (23.8.5), *i.e.*, one needs to solve n_d sets of n_p equations in n_p unknowns of the form:

$$\begin{aligned} 0 = &\sum_{k=1}^{n_\mathrm{p}} \left[\sum_{l} \frac{\partial^3}{\partial x_j \partial x_k \partial x_l} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) \, \frac{\partial}{\partial y_i} \hat{x}_l(\boldsymbol{y}) + 2 \frac{\partial^3}{\partial x_j \partial x_k \partial y_i} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) \right] \frac{\partial}{\partial y_i} \hat{x}_k(\boldsymbol{y}) \\ &+ \sum_{k=1}^{n_\mathrm{p}} \frac{\partial^2}{\partial x_j \, \partial x_k} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) \, \frac{\partial^2}{\partial y_i^2} \hat{x}_k(\boldsymbol{y}) + \frac{\partial^3}{\partial x_j \partial y_i^2} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}), \end{aligned}$$

for $j = 1, \ldots, n_\mathrm{p}$ and $i = 1, \ldots, n_\mathrm{d}$.

23.8.2.2 Scalar parameter

If $n_{\mathrm{p}} = 1$, *i.e.*, $\boldsymbol{x}$ is a scalar, then (23.8.5) simplifies to

$$0 = \left[\frac{\partial^3}{\partial \theta^3} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) \frac{\partial}{\partial y_m} \hat{\boldsymbol{x}}(\boldsymbol{y}) + \frac{\partial^3}{\partial \theta^2 \partial y_m} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) \right] \frac{\partial}{\partial y_i} \hat{\boldsymbol{x}}(\boldsymbol{y})$$
$$+ \frac{\partial^2}{\partial \theta^2} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) \frac{\partial^2}{\partial y_i \, \partial y_m} \hat{\boldsymbol{x}}(\boldsymbol{y}) + \frac{\partial^3}{\partial \theta^2 \partial y_i} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}) \frac{\partial}{\partial y_m} \hat{\boldsymbol{x}}(\boldsymbol{y})$$
$$+ \frac{\partial^3}{\partial \theta \partial y_i \partial y_m} \Psi(\hat{\boldsymbol{x}}(\boldsymbol{y}), \boldsymbol{y}), \; i = 1, \ldots, n_{\mathrm{d}}, \; m = 1, \ldots, n_{\mathrm{d}}. \tag{23.8.7}$$

e,mav,mean,chain2,scalar

By rearranging we can solve explicitly for the second partials of $\hat{\boldsymbol{x}}(\boldsymbol{y})$:

$$\frac{\partial^2}{\partial y_i \, \partial y_m} \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) = \left[-\frac{\partial^2}{\partial \theta^2} \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \right]^{-1}$$
$$\left(\left[\frac{\partial^3}{\partial \theta^3} \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \frac{\partial}{\partial y_m} \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) + \frac{\partial^3}{\partial \theta^2 \partial y_m} \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \right] \frac{\partial}{\partial y_i} \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) \right.$$
$$\left. + \frac{\partial^3}{\partial \theta^2 \partial y_i} \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \frac{\partial}{\partial y_m} \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) + \frac{\partial^3}{\partial \theta \partial y_i \partial y_m} \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}) \right).$$

Substituting this expression into (23.8.3) yields the approximate mean for a scalar parameter estimator.

23.9 Example: regularized least squares (s,mav,rls)

s,mav,rls

The approximations for mean and covariance derived above are exact in the special case where the estimator is linear, because in that case the first-order Taylor expansion (23.3.13) is exact. In this section we verify this property by computing the covariance approximation (23.3.12) and the mean approximation (23.8.5) for a regularized least-squares problem. The expressions are useful for comparing with the corresponding approximations for nonlinear estimators.

Suppose the measurements obey the standard linear model with additive noise:

$$\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon},$$

where $\boldsymbol{A}$ is a known $n_{\mathrm{d}} \times n_{\mathrm{p}}$ matrix. For such problems, the following regularized weighted least-squares cost function is often used for estimation:

$$\Psi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{2} (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x})' \boldsymbol{W} (\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}) + \beta \, \mathsf{R}(\boldsymbol{x}),$$

where $\boldsymbol{W}$ is a positive-semidefinite weighting matrix and $\mathsf{R}(\boldsymbol{x})$ is a roughness penalty of the form

$$\mathsf{R}(\boldsymbol{x}) = \sum_k \psi_k([\boldsymbol{C}\boldsymbol{x}]_k) \,. \tag{23.9.1}$$

e,mav,rls,r

Note that $\nabla^{[1,1]} R(\boldsymbol{x}) = \boldsymbol{0}$, and define

$$\mathbf{R}(\boldsymbol{x}) = \nabla^2 \mathsf{R}(\boldsymbol{x}) = \boldsymbol{C}' \, \mathsf{diag}\left\{ \ddot{\psi}_k([\boldsymbol{C}\boldsymbol{x}]_k) \right\} \boldsymbol{C} \tag{23.9.2}$$

e,mav,rls,r2

to be the matrix of second partials of $\mathsf{R}(\boldsymbol{x})$, where $\ddot{\psi}$ denotes the second derivative of ψ.

Consider the quadratic case where $\psi(x) = x^2/2$, so $\mathbf{R} = \boldsymbol{C}'\boldsymbol{C}$ and assume $\boldsymbol{W}\boldsymbol{A}$ and $\boldsymbol{C}$ have disjoint null spaces. In this case one can derive an explicit expression for the estimator:

$$\hat{\boldsymbol{x}} = \left[\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A} + \beta \mathbf{R} \right]^{-1} \boldsymbol{A}'\boldsymbol{W}\boldsymbol{y}, \tag{23.9.3}$$

e,mav,rls,xh

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" (23.3.12) and (23.8.5).

The partial derivatives of Ψ are:

$$\begin{aligned} \nabla^{[2,0]} \Psi &= \boldsymbol{A}'\boldsymbol{W}\boldsymbol{A} + \beta \boldsymbol{R} \\ -\nabla^{[1,1]} \Psi &= \boldsymbol{A}'\boldsymbol{W} \\ \nabla^{[3,0]} \Psi &= \nabla^{[2,1]} \Psi = \nabla^{[1,2]} \Psi = \boldsymbol{0}, \end{aligned} \tag{23.9.4}$$

so substituting into (23.8.5), one finds that $\nabla^2 \hat{\boldsymbol{x}}(\boldsymbol{y}) = 0$. Thus from (23.8.3):

$$\mathsf{E}[\hat{\boldsymbol{x}}] = \hat{\boldsymbol{x}}(\bar{\boldsymbol{y}}) = \left[\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A} + \beta \boldsymbol{R} \right]^{-1} \boldsymbol{A}'\boldsymbol{W} \, \bar{\boldsymbol{y}},$$

which of course is exactly what one would get from (23.9.3). Substituting (23.9.4) into (23.3.12) yields the estimator covariance:

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} = [\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A} + \beta\boldsymbol{R}]^{-1}\,\boldsymbol{A}'\boldsymbol{W}\,\mathsf{Cov}\{\boldsymbol{y}\}\,\boldsymbol{W}\boldsymbol{A}\,[\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A} + \beta\boldsymbol{R}]^{-1}\,,$$

which again agrees with (23.9.3). If the measurement covariance is known, then usually one chooses $\boldsymbol{W} = \mathsf{Cov}\{\boldsymbol{y}\}^{-1}$, in which case

e,mav,rls,cov

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} = [\mathbf{F} + \beta\boldsymbol{R}]^{-1}\,\mathbf{F}\,[\mathbf{F} + \beta\boldsymbol{R}]^{-1}\,, \tag{23.9.5}$$

where $\mathbf{F} = \boldsymbol{A}'\,\mathsf{Cov}\{\boldsymbol{y}\}^{-1}\,\boldsymbol{A}$ is the Fisher information for estimating $\boldsymbol{x}$ from $\boldsymbol{y}$, when the noise has a normal distribution. The covariance approximation (23.3.16) generalizes (23.9.5).

Because the mean and covariance approximations are exact for quadratic cost functions, one might expect the approximation accuracy for a non-quadratic cost function will depend on how far the cost function departs from being quadratic. Many cost 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 (23.3.13) it is clear that as the noise variance goes to zero, we will have $y_i \to \bar{y}_i$, so the Taylor approximation error will vanish. This asymptotic property is illustrated empirically in [4].

s,mav,ex,trans

23.10 Example: transmission tomography (s,mav,ex,trans)

To illustrate the accuracy of the approximation for estimator covariance given by (23.3.12), 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 [39]–[41]. Because PET transmission scans are essentially measurements of nuisance parameters, one would like to use very short transmission scans. Because short scans have fewer counts (lower SNR), the conventional linear filter-backproject (FBP) reconstruction method performs poorly. Statistical methods have the potential to significantly reduce the error variance, but because 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 x_j denotes the attenuation coefficient in the jth pixel. The transmission measurements have independent Poisson distributions, and we assume the mean of y_i is:

$$\begin{aligned} \bar{Y}_i(\boldsymbol{x}) &= T p_i(\boldsymbol{x}) \\ p_i(\boldsymbol{x}) &= b_i e^{-\sum_j a_{ij} x_j} + r_i, \end{aligned} \tag{23.10.1}$$

where the a_{ij} factors denote the intersection length of the nth ray passing though the jth pixel, $\{b_i\}$ denote the rates of emissions from the transmission source, $\{r_i\}$ 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:

eqn-like-trans

$$\mathsf{L}(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n_\mathrm{d}} y_i \log \bar{Y}_i(\boldsymbol{x}) - \bar{Y}_i(\boldsymbol{x}), \tag{23.10.2}$$

neglecting constants independent of $\boldsymbol{x}$. Because tomography is ill-conditioned, rather than performing ordinary ML estimation, many investigators have used penalized-likelihood cost functions of the form[5]

e,mav,cost-trans

$$\Psi(\boldsymbol{x}, \boldsymbol{y}) = \frac{1}{T} \mathsf{L}(\boldsymbol{x}, \boldsymbol{y}) - \beta R(\boldsymbol{x}), \tag{23.10.3}$$

where the roughness penalty R was defined in (23.9.1).

Due to the nonlinearity of (23.10.1) and the non-quadratic likelihood function (23.10.2) for Poisson statistics, the estimate $\hat{\boldsymbol{x}}$ formed by maximizing (23.10.3) is presumably a very nonlinear function of $\boldsymbol{y}$. Furthermore, because attenuation coefficients are nonnegative, one usually enforces the inequality constraint $\hat{\boldsymbol{x}} \succeq \mathbf{0}$. Therefore this problem provides a stringent test of the accuracy of the mean and variance approximations.

23.10.1 Covariance Approximation

Because the number of measurements (or rays) n_d and the number of parameters (pixels) p are both large, we would like to approximate the variance of certain pixels of interest using (23.4.2), which requires the following partial derivatives:

$$\begin{aligned} \frac{\partial}{\partial x_j} \mathsf{L}(\boldsymbol{x}, \boldsymbol{y}) &= T \sum_{i=1}^{n_\mathrm{d}} a_{ij} \left(1 - \frac{y_i}{\bar{Y}_i(\boldsymbol{x})}\right) (p_i(\boldsymbol{x}) - r_i) \\ -\frac{\partial^2}{\partial x_j \, \partial x_k} \mathsf{L}(\boldsymbol{x}, \boldsymbol{y}) &= T \sum_{i=1}^{n_\mathrm{d}} a_{ij} a_{ik} \, q_i(\boldsymbol{x}) \\ q_i(\boldsymbol{x}) &= \left(1 - \frac{r_i y_i / T}{p_i^2(\boldsymbol{x})}\right) b_i e^{-\sum_j a_{ij} x_j} \\ \frac{\partial^2}{\partial x_j \, \partial y_i} \mathsf{L}(\boldsymbol{x}, \boldsymbol{y}) &= -a_{ij} \left(1 - \frac{r_i}{p_i(\boldsymbol{x})}\right). \end{aligned}$$

Combining the above expressions in matrix form with the expressions for the partials of R given in §23.9:

$$\begin{aligned} -\nabla^{20} \Psi(\boldsymbol{x}, \boldsymbol{y}) &= \boldsymbol{A}' \, \mathsf{diag}\{q_i(\boldsymbol{x})\} \, \boldsymbol{A} + \beta \boldsymbol{R}(\boldsymbol{x}) \\ \nabla^{11} \Psi(\boldsymbol{x}, \boldsymbol{y}) &= -\frac{1}{T} \boldsymbol{A}' \, \mathsf{diag}\left\{1 - \frac{r_i}{p_i(\boldsymbol{x})}\right\}, \end{aligned}$$

where $\boldsymbol{A} = \{a_{ij}\}$ is a large sparse matrix, and $\mathsf{diag}\{v_i\}$ denotes a $n_\mathrm{d} \times n_\mathrm{d}$ diagonal matrix with elements $v_1, \ldots, v_{n_\mathrm{d}}$ along the diagonal. For simplicity we focus on the case where $r_i = 0$, in which case $q_i(\boldsymbol{x}) = p_i(\boldsymbol{x})$ and the above expressions simplify to

$$-\nabla^{20} \Psi(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{A}' \, \mathsf{diag}\{p_i(\boldsymbol{x})\} \, \boldsymbol{A} + \beta \boldsymbol{R}(\boldsymbol{x})$$

[5] Due to the $\frac{1}{T}$ term in (23.10.3), one can show that for a fixed β, as $T \to \infty$, the maximum penalized-likelihood estimate $\hat{\boldsymbol{x}}$ will converge in probability to $\tilde{\boldsymbol{x}}$, a biased estimate [6]. For asymptotically unbiased estimates, one must let $\beta \to 0$ at an appropriate rate as $T \to \infty$ [11].

$$\nabla^{11}\,\Psi(\boldsymbol{x},\boldsymbol{y}) \quad = \quad -\frac{1}{T}\boldsymbol{A}'.$$

By the assumption that the measurements have independent Poisson distributions, $\mathsf{Cov}\{\boldsymbol{y}\} = \mathsf{diag}\{\bar{Y}_i(\boldsymbol{x}_{\rm true})\}$. Substituting into (23.3.12) and simplifying yields the following approximation to the estimator covariance:

eqn-cov-trans

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \approx \frac{1}{T}\left[\mathsf{F}(\check{\boldsymbol{x}}) + \beta\boldsymbol{R}(\check{\boldsymbol{x}})\right]^{-1}\mathsf{F}(\boldsymbol{x}_{\rm true})\left[\mathsf{F}(\check{\boldsymbol{x}}) + \beta\boldsymbol{R}(\check{\boldsymbol{x}})\right]^{-1}, \tag{23.10.4}$$

where

eqn-f

$$\mathsf{F}(\boldsymbol{x}) = \boldsymbol{A}'\,\mathsf{diag}\{p_i(\boldsymbol{x})\}\,\boldsymbol{A} \tag{23.10.5}$$

is $1/T$ times the Fisher information for estimating $\boldsymbol{x}$ from $\boldsymbol{y}$. Note the similarity to (23.9.5).

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

- Compute $\check{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \Psi(\boldsymbol{x},\bar{\boldsymbol{y}})$ by applying to noise-free data $\bar{\boldsymbol{y}}$ a maximization algorithm such as the fast converging coordinate-ascent algorithm of Bouman and Sauer [42], [43].
- Forward project $\check{\boldsymbol{x}}$ to compute $p_i(\check{\boldsymbol{x}}) = \sum_j a_{ij}\check{\boldsymbol{x}}_j + r_i$. Likewise for $p_i(\boldsymbol{x}_{\rm true})$.
- Pick a pixel j of interest and solve the equation $\left[\boldsymbol{A}'\,\mathsf{diag}\{p_i(\check{\boldsymbol{x}})\}\,\boldsymbol{A} + \beta\boldsymbol{R}(\check{\boldsymbol{x}})\right]\boldsymbol{u}_j = \boldsymbol{e}_j$ for $\boldsymbol{u}_j$ using a fast iterative method such as preconditioned conjugate gradients [44] or Gauss-Siedel [31].
- Compute $\frac{1}{T}(\boldsymbol{u}_j)'\boldsymbol{A}'\,\mathsf{diag}\{p_i(\boldsymbol{x}_{\rm true})\}\,\boldsymbol{A}\,\boldsymbol{u}_j$ by first forward projecting $\boldsymbol{u}_j$ to compute $v = \boldsymbol{A}\,\boldsymbol{u}_j$, and then summing:
$$\mathsf{Var}\{\hat{x}_j\} \approx \frac{1}{T}\sum_{i=1}^{n_{\rm d}} v_i^2 p_i(\boldsymbol{x}_{\rm true}).$$

The overall computational requirements for this recipe are roughly equivalent to two maximizations of Ψ. 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 Ψ.

23.10.2 Empirical Results

To assess the accuracy of approximation (23.10.4), we performed numerical simulations using the synthetic attenuation map shown in Fig. 23.14.1 as $\boldsymbol{x}_{\rm true}$. 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_{ij} 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_i 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 $\sum_{i=1}^{n_{\rm d}} \bar{Y}_i(\boldsymbol{x}_{\rm true})$ was 0.25, 1, 4, and 16 million counts. We set $r_i = 0$ for simplicity. For each study, we generated 100 realizations of pseudo-random Poisson transmission measurements according to (23.10.1) and then reconstructed using the penalized-likelihood estimator described by (23.10.3) using a coordinate-ascent algorithm [41]. This algorithm enforced the nonnegativity constraint $\hat{\boldsymbol{x}} \succeq \boldsymbol{0}$. For simplicity, we used the function $\psi(x) = x^2/2$ for the penalty in (23.9.1). 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 (23.10.4). Fig. 23.14.2 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 (23.10.4) 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 (23.8.1) is reasonably accurate for maximum-likelihood estimators [20], [21], [38]; we have found similar results for penalized-likelihood estimators such as (23.10.3). (This is fortuitous because the 2nd-order expressions for mean are considerably more expensive to compute because $p = 128 \cdot 64$ and $n_{\rm d} = 192 \cdot 96$ are very large in this example.) Fig. 23.14.3 displays a representative cross-section through the mean predicted by (23.8.1) 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 (23.8.1) and (23.3.12) are useful for predicting penalized-likelihood estimator performance in transmission tomography.

23.11 Post-estimation plug-in variance approximation

s,mav,plug

The approximation (23.3.12) for the estimator covariance depends on both $\check{\boldsymbol{x}}$ and $\mathsf{Cov}\{\boldsymbol{y}\}$, so as written its primary use will be in computer simulations where $\check{\boldsymbol{x}}$ and $\mathsf{Cov}\{\boldsymbol{y}\}$ 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 $\boldsymbol{x}_{\text{true}}$ is unknown, and $\mathsf{Cov}\{\boldsymbol{y}\}$ may also be unknown. In some problems this can be done using a "plug-in" estimate in which we substitute the estimate $\hat{\boldsymbol{x}}$ in for $\check{\boldsymbol{x}}$ in (23.3.12). The effectiveness of this approach will undoubtedly 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 $\boldsymbol{y}$ and a penalized-likelihood estimate $\hat{\boldsymbol{x}}$ computed by maximizing the cost function (23.10.3). If we knew $\check{\boldsymbol{x}}$ and $\boldsymbol{x}_{\text{true}}$, then we could use (23.10.4) to approximate the covariance of $\hat{\boldsymbol{x}}$. If we only have $\hat{\boldsymbol{x}}$, then in light of the form of the covariance approximation given by (23.10.4), a natural approach to estimating the covariance would be to simply plug-in $\hat{\boldsymbol{x}}$ for $\check{\boldsymbol{x}}$ and $\boldsymbol{x}_{\text{true}}$ in (23.10.4):

$$\widehat{\mathsf{Cov}\{\hat{\boldsymbol{x}}\}} = \frac{1}{T}\left[\mathsf{F}(\hat{\boldsymbol{x}}) + \beta \boldsymbol{R}(\hat{\boldsymbol{x}})\right]^{-1} \mathsf{F}(\hat{\boldsymbol{x}}) \left[\mathsf{F}(\hat{\boldsymbol{x}}) + \beta \boldsymbol{R}(\hat{\boldsymbol{x}})\right]^{-1},$$

from which one can compute estimates of the variance of individual pixels or region-of-interest values using the same technique as in (23.4.2).

At first it may seem unlikely that such a simplistic approach would yield reliable estimates of variability. However, note that in the definition (23.10.5) of $\mathsf{F}(\boldsymbol{x})$, the *only* dependence on $\boldsymbol{x}$ is through its *projections* $p_i(\boldsymbol{x})$. 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 $\hat{\boldsymbol{x}}$ agree reasonably well with $\check{\boldsymbol{x}}$ and $\boldsymbol{x}_{\text{true}}$, then the projections $p_i(\hat{\boldsymbol{x}})$, $p_i(\check{\boldsymbol{x}})$, and $p_i(\boldsymbol{x}_{\text{true}})$ will be very similar. Furthermore, the dependence on the p_i terms in (23.10.4) is through a diagonal matrix that is sandwiched between the $\boldsymbol{A}'$ and $\boldsymbol{A}$ 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. 23.14.1. If $\hat{\boldsymbol{x}}^{(m)}$ denotes the mth realization $(m = 1, \ldots, 100)$, then the mth estimate of the standard deviation of $\hat{x}_j$ is:

$$\hat{\sigma}_j^{(m)} = \left[(\boldsymbol{e}_j)' \widehat{\mathsf{Cov}\{\hat{\boldsymbol{x}}\}} \boldsymbol{e}_j\right]^{1/2}$$

$$= \left[(\boldsymbol{e}_j)' \frac{1}{T}\left[\mathsf{F}(\hat{\boldsymbol{x}}) + \beta \boldsymbol{R}(\hat{\boldsymbol{x}})\right]^{-1} \mathsf{F}(\hat{\boldsymbol{x}}) \left[\mathsf{F}(\hat{\boldsymbol{x}}) + \beta \boldsymbol{R}(\hat{\boldsymbol{x}})\right]^{-1} \boldsymbol{e}_j\right]^{1/2}. \tag{23.11.1}$$

eqn-plug-var-trans

Histograms of the standard deviation estimates $\left\{\hat{\sigma}_j^{(m)}\right\}_{m=1}^{100}$ are shown in Fig. 23.14.4 and Fig. 23.14.5 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 [45].

23.12 Example: Emission Tomography

s,mav,ex,emis

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 [39], [46].

In emission tomography the parameter x_j denotes the radionuclide concentration in the jth pixel. The emission measurements have independent Poisson distributions, and we assume the mean of y_i is:

$$\begin{aligned} \bar{Y}_i(\boldsymbol{x}) &= T p_i(\boldsymbol{x}) \\ p_i(\boldsymbol{x}) &= \sum_j a_{ij} x_j + r_i, \end{aligned} \tag{23.12.1}$$

where the a_{ij} are proportional to the probability that an emission in voxel j is detected by the nth detector pair, $\{r_i\}$ 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 (23.10.2), but with definition (23.12.1) for $\bar{Y}_i(\boldsymbol{x})$. We again focus on penalized-likelihood cost functions of the form (23.10.3).

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.

23.12.1 Covariance Approximation

Approximating the variance of certain pixels of interest using (23.4.2) requires the following partial derivatives:

$$\begin{aligned}
\frac{\partial}{\partial x_j} \mathsf{L}(\boldsymbol{x}, \boldsymbol{y}) &= T \sum_{i=1}^{n_\mathrm{d}} a_{ij} \left(\frac{y_i}{\bar{Y}_i(\boldsymbol{x})} - 1 \right) \\
-\frac{\partial^2}{\partial x_j \, \partial x_k} \mathsf{L}(\boldsymbol{x}, \boldsymbol{y}) &= T \sum_{i=1}^{n_\mathrm{d}} a_{ij} a_{ik} \frac{y_i/T}{p_i^2(\boldsymbol{x})} \\
\frac{\partial^2}{\partial x_j \, \partial y_i} \mathsf{L}(\boldsymbol{x}, \boldsymbol{y}) &= a_{ij}/p_i(\boldsymbol{x}).
\end{aligned}$$

Combining the above expressions in matrix form with the expressions for the partials of R given in §23.9:

$$\begin{aligned}
-\nabla^{20} \Psi(\boldsymbol{x}, \boldsymbol{y}) &= \boldsymbol{A}' \, \mathsf{diag}\left\{ \frac{y_i/T}{p_i^2(\boldsymbol{x})} \right\} \boldsymbol{A} + \beta \boldsymbol{R}(\boldsymbol{x}) \\
\nabla^{11} \Psi(\boldsymbol{x}, \boldsymbol{y}) &= -\frac{1}{T} \boldsymbol{A}' \, \mathsf{diag}\left\{ \frac{1}{p_i(\boldsymbol{x})} \right\}.
\end{aligned}$$

Thus

$$\begin{aligned}
-\nabla^{20} \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}) &= \boldsymbol{A}' \, \mathsf{diag}\left\{ \frac{p_i(\boldsymbol{x}_\mathrm{true})}{p_i^2(\check{\boldsymbol{x}})} \right\} \boldsymbol{A} + \beta \boldsymbol{R}(\check{\boldsymbol{x}}) \\
\nabla^{11} \Psi(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}) &= -\frac{1}{T} \boldsymbol{A}' \, \mathsf{diag}\left\{ \frac{1}{p_i(\check{\boldsymbol{x}})} \right\}.
\end{aligned}$$

By the assumption that the measurements have independent Poisson distributions, $\mathsf{Cov}\{\boldsymbol{y}\} = \mathsf{diag}\{\bar{Y}_i(\boldsymbol{x}_\mathrm{true})\}$. Substituting into (23.3.12) and simplifying yields the following approximation to the estimator covariance:

eqn-cov-emis

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \approx \frac{1}{T} \left[\mathsf{F} + \beta \boldsymbol{R}(\check{\boldsymbol{x}}) \right]^{-1} \mathsf{F} \left[\mathsf{F} + \beta \boldsymbol{R}(\check{\boldsymbol{x}}) \right]^{-1}, \tag{23.12.2}$$

where

eqn-f-trans

$$\mathsf{F} = \boldsymbol{A}' \, \mathsf{diag}\left\{ \frac{p_i(\boldsymbol{x}_\mathrm{true})}{p_i^2(\check{\boldsymbol{x}})} \right\} \boldsymbol{A}. \tag{23.12.3}$$

We compute the approximate variance of $\hat{x}_j$ using a recipe similar to that given in §23.10.

23.12.2 Empirical Results

To assess the accuracy of approximation (23.12.2), we performed numerical simulations using the synthetic brain image shown in Fig. 23.14.6 as $\boldsymbol{x}_\mathrm{true}$, 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_{ij} 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 $\sum_{i=1}^{n_\mathrm{d}} \bar{Y}_i(\boldsymbol{x}_\mathrm{true})$ was 0.2, 0.8, 3.2, and 12.8 million counts. The r_i 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 (23.12.1) and then reconstructed using a space-alternating generalized EM algorithm [46], that enforces the nonnegativity constraint $\hat{\boldsymbol{x}} \succeq \mathbf{0}$. FBP images served as the initial estimate for the iterative algorithm.

For the penalty function ψ we studied two cases: the simple quadratic case $\psi(x) = x^2/2$, as well as a nonquadratic penalty: the third entry in Table III of [47]:

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

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 (23.10.4) 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 Fig. 23.14.7 and Fig. 23.14.8. 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 Fig. 23.14.9 and Fig. 23.14.10. 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 (23.3.12), (23.12.2) gives reasonably accurate predictions of estimator performance, with better agreement at higher SNR.

We also investigated the post-estimation plug-in approach described in §23.11 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 $\hat{\boldsymbol{x}}$ to (23.12.2) yields essentially the same value as one gets by using $\check{\boldsymbol{x}}$ and $\boldsymbol{x}_{\text{true}}$. Thus it appears that the intrinsic error in the approximation (23.12.2) is more significant than the differences between $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}_{\text{true}}$. 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 $\boldsymbol{x}_{\text{true}}$ is unknown.

As illustrated by Fig. 23.14.11, the 0th-order mean approximation (23.8.1) 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.

23.12.3 Mean: 2nd Order

This subsection illustrates an application of the second-order approximation for estimator mean given by (23.8.6). In the routine practice of PET and SPECT, images are reconstructed using non-statistical Fourier methods [48]. Often one can obtain more accurate images using likelihood-based methods. Because 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 with a weighted least-squares or *quadratic* cost function for which there are iterative algorithms that converge faster (*e.g.*, [42], [43], [49], [50]). Unfortunately, in the context of *transmission* tomography, quadratic cost functions lead to estimation *bias* for low-count measurements [41]. To determine whether a similar undesirable bias exists for the quadratic approximation in the *emission* case, we now use the analytical expression (23.8.6) 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 [42], [43], the quadratic approximation to the log-likelihood $\Psi_L(\boldsymbol{x}, Y) = \mathsf{L}(\boldsymbol{x}, Y)$ is

$$\Psi_Q(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{2} \sum_{n \,:\, y_i > 0} \frac{1}{y_i} (y_i - \bar{Y}_i(\boldsymbol{x}))^2.$$

The cost functions Ψ_L and Ψ_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_i/a_i is a constant independent of n.

For large images, the computation required for solving (23.8.6) 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 23.1 we derive the partial derivatives necessary for evaluating (23.8.6) for each cost function (for $p = 1$). In this table $\mathsf{F}_{\boldsymbol{x}}$ denotes the Fisher information for estimating $\boldsymbol{x}$ from $\boldsymbol{y}$:

$$\mathsf{F}_{\boldsymbol{x}} = -\,\mathsf{E}\big[\nabla^2_{\boldsymbol{x}} \log f(\boldsymbol{y}, \boldsymbol{x})\big] = \sum_{i=1}^{n_\text{d}} a_i^2 / \bar{Y}_i(\boldsymbol{x}) = \sum_{i=1}^{n_\text{d}} \frac{a_i^2}{a_i \theta + r_i}.$$

The second and final two rows of Table 23.1 show three important points:

- For each cost function, $\nabla^{10}\, \Psi(\boldsymbol{x}, \bar{\boldsymbol{y}}(\boldsymbol{x})) = 0$, so that $\check{\boldsymbol{x}} = h(\bar{\boldsymbol{y}}(\boldsymbol{x})) = \boldsymbol{x}$, *i.e.*, the estimators work perfectly with noiseless data. Therefore the 0th-order approximation (23.8.1) yields $\mathsf{E}[\hat{\boldsymbol{x}}] = \boldsymbol{x}$, which is inaccurate for the Ψ_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. 23.14.12 compares the bias predicted analytically using the approximation (23.8.6) with an empirically computed bias performed by numerical simulations. In these simulations we used $\boldsymbol{x}_{\text{true}} = 1$, $r_i = 0$, $a_i = 1$, and $n_\text{d} = 10$, and varied T so that $\frac{1}{n_\text{d}} \sum_{i=1}^{n_\text{d}} \bar{Y}_i(\boldsymbol{x}_{\text{true}})$ (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_i. When the bias due to the quadratic

approximation is significant, one can apply a hybrid Poisson/polynomial cost function similar to that proposed for transmission tomography [41]. 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 cost function without introducing bias [41].

23.13 Discussion

We have derived approximations for the mean and covariance of estimators that are defined as the maximum of some cost 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 cost function subject to certain equality constraints. One can use methods similar to the derivation of the constrained *Cramér-Rao* lower bound [51], [52] to generalize the covariance approximation (23.3.12) 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 §23.10. In the future, we expect to evaluate the post-reconstruction estimate of region variability §23.11 for performing weighted estimates of kinetic parameters from dynamic PET emission scans [45]. 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. Because FBP is a linear reconstruction algorithm, it is straightforward to compute estimates of variability for Poisson emission measurements [45], [53]. 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.

23.14 Appendix

This appendix synopsizes the asymptotic variance of M-estimates given by Serfling [6]. The results in Serfling are for a scalar parameter $\boldsymbol{x}$, so we consider the scalar case below. (See [27] for the multiparameter case.) As in §(I), let $\Phi(\boldsymbol{x}, Y)$ be the cost function that is to be maximized to find $\hat{\boldsymbol{x}}$, and define

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

Assume Y has a probability distribution $F(y; \boldsymbol{x}_{\text{true}})$, and let $\bar{\boldsymbol{x}}$ be the value of $\boldsymbol{x}$ that satisfies:

$$\int \psi(\boldsymbol{x}, y) \, dF(y; \boldsymbol{x}_{\text{true}}) = 0. \tag{23.14.1}$$

Serfling [6] shows that $\hat{\boldsymbol{x}}$ is asymptotically normal with mean $\bar{\boldsymbol{x}}$ and variance

$$\frac{\int \psi^2(\bar{\boldsymbol{x}}, y) \, dF(y; \boldsymbol{x}_{\text{true}})}{\left[\frac{\partial}{\partial \boldsymbol{x}} \int \psi^2(\boldsymbol{x}, y) \, dF(y; \boldsymbol{x}_{\text{true}}) \Big|_{\boldsymbol{x} = \bar{\boldsymbol{x}}} \right]^2}. \tag{23.14.2}$$

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

- The term $\bar{\boldsymbol{x}}$ plays a role similar to our $\check{\boldsymbol{x}}$, but solving the integral equation (23.14.1) for $\bar{\boldsymbol{x}}$ is in general more work than calculating $\check{\boldsymbol{x}}$ by maximizing $\Phi(\cdot, \bar{\boldsymbol{y}})$.
- Both $\bar{\boldsymbol{x}}$ and the expression for the asymptotic variance depend on the entire measurement distribution $F(y; \boldsymbol{x}_{\text{true}})$, whereas our approximation depends only on the mean and covariance of the measurements.

With some additional work, one can show that if $\psi(\boldsymbol{x}, Y)$ is affine in Y, then $\bar{\boldsymbol{x}}$ and $\check{\boldsymbol{x}}$ are equal, and (23.14.2) is equivalent to (23.3.12). Both gaussian and Poisson measurements yield ψ that are affine in Y (cf (23.10.2)), so (23.3.12) is the asymptotic covariance in those cases, provided the penalty is data-independent. For data-dependent penalties [54] or for more complicated noise distributions, such as the Poisson/gaussian model for CCD arrays [55], the covariance approximation given by (23.3.12) will probably be easier to implement than (23.14.2).

	Objective	
Term	Likelihood	Quadratic
$\Psi(\boldsymbol{x}, \boldsymbol{y})$	$\sum_{i=1}^{n_\mathrm{d}} y_i \log \bar{Y}_i(\boldsymbol{x}) - \bar{Y}_i(\boldsymbol{x})$	$-\frac{1}{2} \sum_{i=1}^{n_\mathrm{d}} (y_i - \bar{Y}_i(\boldsymbol{x}))^2 / y_i$
$\frac{\partial}{\partial \theta} \Psi(\boldsymbol{x}, \boldsymbol{y})$	$\sum_{i=1}^{n_\mathrm{d}} a_i (y_i / \bar{Y}_i(\boldsymbol{x}) - 1)$	$\sum_{i=1}^{n_\mathrm{d}} a_i (1 - \bar{Y}_i(\boldsymbol{x}) / y_i)$
$-\frac{\partial^2}{\partial \theta^2} \Psi(\boldsymbol{x}, \boldsymbol{y})$	$\sum_{i=1}^{n_\mathrm{d}} a_i^2 y_i / \bar{Y}_i(\boldsymbol{x})^2$	$\sum_{i=1}^{n_\mathrm{d}} a_i^2 / y_i$
$\frac{\partial^3}{\partial \theta^3} \Psi(\boldsymbol{x}, \boldsymbol{y})$	$\sum_{i=1}^{n_\mathrm{d}} 2 a_i^3 y_i / \bar{Y}_i(\boldsymbol{x})^3$	0
$\frac{\partial^2}{\partial \theta \, \partial y_i} \Psi(\boldsymbol{x}, \boldsymbol{y})$	$a_i / \bar{Y}_i(\boldsymbol{x})$	$a_i \bar{Y}_i(\boldsymbol{x}) / y_i^2$
$\frac{\partial^3}{\partial \theta \partial y_i^2} \Psi(\boldsymbol{x}, \boldsymbol{y})$	0	$-2 a_i \bar{Y}_i(\boldsymbol{x}) / y_i^3$
$\frac{\partial^3}{\partial \theta^2 \partial y_i} \Psi(\boldsymbol{x}, \boldsymbol{y})$	$-a_i^2 / \bar{Y}_i(\boldsymbol{x})^2$	a_i^2 / y_i^2
$-\frac{\partial^2}{\partial \theta^2} \Psi(\boldsymbol{x}, \bar{\boldsymbol{y}})$	$\mathsf{F}_{\boldsymbol{x}}$	$\mathsf{F}_{\boldsymbol{x}}$
$\frac{\partial^3}{\partial \theta^3} \Psi(\boldsymbol{x}, \bar{\boldsymbol{y}})$	$2 \sum_{i=1}^{n_\mathrm{d}} a_i^3 / \bar{Y}_i^2$	0
$\frac{\partial^2}{\partial \theta \, \partial y_i} \Psi(\boldsymbol{x}, \bar{\boldsymbol{y}})$	$a_i / \bar{Y}_i$	$a_i / \bar{Y}_i$
$\frac{\partial^3}{\partial \theta \partial y_i^2} \Psi(\boldsymbol{x}, \bar{\boldsymbol{y}})$	0	$-2 a_i / \bar{Y}_i^2$
$\frac{\partial^3}{\partial \theta^2 \partial y_i} \Psi(\boldsymbol{x}, \bar{\boldsymbol{y}})$	$-a_i^2 / \bar{Y}_i^2$	$a_i^2 / \bar{Y}_i^2$
$\frac{\partial}{\partial y_i} h(\bar{\boldsymbol{y}})$	$a_i / (\bar{Y}_i \mathsf{F}_{\boldsymbol{x}})$	$a_i / (\bar{Y}_i \mathsf{F}_{\boldsymbol{x}})$
$\frac{\partial^2}{\partial y_i^2} h(\bar{\boldsymbol{y}})$	$\frac{2}{\mathsf{F}_{\boldsymbol{x}}^2} \frac{a_i^2}{\bar{Y}_i^2} \left[\frac{1}{\mathsf{F}_{\boldsymbol{x}}} \sum_{i=1}^{n_\mathrm{d}} a_i^3 / \bar{Y}_i^2 - a_i / \bar{Y}_i \right]$	$\frac{2}{\mathsf{F}_{\boldsymbol{x}}} \frac{a_i}{\bar{Y}_i^2} \left[\frac{a_i^2}{\bar{Y}_i \mathsf{F}_{\boldsymbol{x}}} - 1 \right]$
$\mathsf{Var}\{\hat{\boldsymbol{x}}\} \approx$	$1 / \mathsf{F}_{\boldsymbol{x}}$	$1 / \mathsf{F}_{\boldsymbol{x}}$
$\mathsf{E}[\hat{\boldsymbol{x}}] - \boldsymbol{x} \approx$	0	$\frac{1}{\mathsf{F}_{\boldsymbol{x}}^2} \sum_{i=1}^{n_\mathrm{d}} \frac{a_i^3}{\bar{Y}_i^2} - \frac{1}{\mathsf{F}_{\boldsymbol{x}}} \sum_{i=1}^{n_\mathrm{d}} \frac{a_i}{\bar{Y}_i}$

table-emis

Table 23.1: Objective functions and partial derivatives for scalar emission tomography problem with $\bar{Y}_i(\boldsymbol{x}) = a_i \boldsymbol{x}$.

Figure 23.14.1: Simulated thorax attenuation map used to evaluate the mean and variance approximations for penalized-likelihood estimators in transmission tomography. fig-object-trans

Figure 23.14.2: Variance for center pixel of attenuation map as predicted by (23.10.4) compared with simulation results from penalized-likelihood estimator (23.10.3). Also shown is the variance of conventional FBP. fig-std

Figure 23.14.3: 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. fig-mean

Figure 23.14.4: Histogram of 100 post-reconstruction plug-in estimates of variability $\mathsf{Var}\{\hat{x}_j\}$ described by (23.11.1), where j corresponds to the center pixel of the attenuation map shown in Fig. 23.14.1, for 250K count measurements. The empirical standard deviation from 100 realizations was $1.74 \cdot 10^{-3}\mathrm{mm}^{-1}$. fig-hist-trans-2p5

Figure 23.14.5: As in previous figure, but for 1M count measurements. The empirical standard deviation from 100 realizations was $9.30 \cdot 10^{-4}\mathrm{mm}^{-1}$. fig-hist-trans-10

fig,obj,emis

Figure 23.14.6: Simulated brain radioisotope emission distribution.

Figure 23.14.7: Comparison of predicted variance from (23.12.2) with empirical performance of penalized-likelihood emission image reconstruction with quadratic penalty for pixel at center of thalamus. fig,std,49,64,-2,quad,1,-,8e5,r10

fig,std,53,67,-2,quad,1,-,8e5,r10

Figure 23.14.8: As in Fig. 23.14.7 but for pixel at edge of thalamus.

fig,std,49,64,-2,lange3,1,-,1p0,3,8e5,r10

Figure 23.14.9: As in Fig. 23.14.7 but for nonquadratic penalty (see text).

fig,std,53,67,-2,lange3,1,-,1p0,3,8e5,r10

Figure 23.14.10: As in Fig. 23.14.8 but for nonquadratic penalty (see text).

Figure 23.14.11: Horizontal profile through emission phantom, 0th-order predicted mean, and empirical mean from penalized-likelihood estimator using nonquadratic penalty for 0.8M count case. fig,mean,emis,lange3

Figure 23.14.12: 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. fig-emis-bias

s,mav,pl,asymp

23.15 Asymptotics of penalized-likelihood estimators (s,mav,pl,asymp)

It may be useful to consider when and how the covariance approximations in §23.3 become more accurate "asymptotically." To address this, we must first consider what are the appropriate asymptotics for imaging problems. One might like to consider cases where we collect more and more projection views in tomography, or more and more k-space data in MRI, but those types of asymptotics are not considered here. Instead, we consider the hypothetical case where we acquire repeated scans of the same (unchanging) object and imagine reconstructing a single image from N such scans and consider what happens as N increases. For Poisson statistics, this turns out to be equivalent to letting the number of collected counts increase asymptotically.

If we were to collect N repeated scans, then it would be reasonable to consider the following penalized-likelihood cost function:

$$\Psi(\boldsymbol{x}, \boldsymbol{y}) = \sum_{n=1}^{N} \text{Ł}(\boldsymbol{x}, \boldsymbol{y}_n) + \beta_N R(\boldsymbol{x}),$$

where Ł denotes the negative log-likelihood, and $\{\boldsymbol{y}_n\}$ are i.i.d. random vectors.

Assume that $\beta_N / N \to \beta \geq 0$ as $N \to \infty$, and define

$$\check{\boldsymbol{x}} \triangleq \arg\min_{\boldsymbol{x}} \; \mathsf{E}[\text{Ł}(\boldsymbol{x}, \boldsymbol{y}_1)] + \beta R(\boldsymbol{x}).$$

In other words, $\check{\boldsymbol{x}}$ denotes the estimate computed from the average penalized-likelihood function, *cf.* [27, p. 132]. For certain statistical models, such as the normal and Poisson cases, the log-likelihood is effectively affine in the data, in which case

$$\mathsf{E}[\text{Ł}(\boldsymbol{x}, \boldsymbol{y}_1)] = \text{Ł}(\boldsymbol{x}, \bar{\boldsymbol{y}}_1),$$

where $\bar{\boldsymbol{y}}_n = \mathsf{E}[\boldsymbol{y}_n]$. In such cases, $\check{\boldsymbol{x}}$ is interpreted as the estimate given noiseless data. However, we need not use this special case in what follows. Define the negative log-likelihood Hessian $\boldsymbol{L}$ by

$$\boldsymbol{L} \triangleq \mathsf{E}\left[\nabla^{[2,0]} \text{Ł}(\check{\boldsymbol{x}}, \boldsymbol{y}_1)\right],$$

and let $\mathbf{R}$ denote the Hessian of $\mathsf{R}(\boldsymbol{x})$ at $\check{\boldsymbol{x}}$.

Because the $\boldsymbol{y}_n$ vectors are independent,

$$\mathsf{Cov}\{\boldsymbol{\Gamma}(\check{\boldsymbol{x}}, \boldsymbol{y})\} = N \, \mathsf{Cov}\{\nabla_{\boldsymbol{x}} \text{Ł}(\check{\boldsymbol{x}}, \boldsymbol{y}_1)\},$$

where $\boldsymbol{\Gamma}$ was defined in (23.3.1). Thus the covariance approximation (23.3.5) simplifies to

$$\begin{aligned} \mathsf{Cov}\{\hat{\boldsymbol{x}}\} &\approx [N\boldsymbol{L} + \beta_N \mathbf{R}]^{-1} \, N \, \mathsf{Cov}\{\nabla_{\boldsymbol{x}} \text{Ł}(\check{\boldsymbol{x}}, \boldsymbol{y}_1)\} \, [N\boldsymbol{L} + \beta_N \mathbf{R}]^{-1} \\ &= \frac{1}{N} \left[\boldsymbol{L} + \frac{\beta_N}{N} \mathbf{R}\right]^{-1} \mathsf{Cov}\{\nabla_{\boldsymbol{x}} \text{Ł}(\check{\boldsymbol{x}}, \boldsymbol{y}_1)\} \left[\boldsymbol{L} + \frac{\beta_N}{N} \mathbf{R}\right]^{-1}. \end{aligned}$$

The form of this covariance approximation suggests the following asymptotic result, which is shown rigorously in [27, p. 133] [56] Under reasonable regularity conditions on the likelihood Ł and penalty function $\boldsymbol{R}$, the normalized error vector

$$\sqrt{N}(\hat{\boldsymbol{x}} - \check{\boldsymbol{x}})$$

is *asymptotically normally distributed* with mean zero and covariance

e,mav,cov,pl

$$[\boldsymbol{L} + \beta \mathbf{R}]^{-1} \, \mathsf{Cov}\{\nabla_{\boldsymbol{x}} \text{Ł}(\check{\boldsymbol{x}}, \boldsymbol{y}_1)\} \, [\boldsymbol{L} + \beta \mathbf{R}]^{-1}. \tag{23.15.1}$$

Furthermore, if Ł is effectively affine in $\boldsymbol{y}_n$, then

$$\mathsf{Cov}\{\nabla_{\boldsymbol{x}} \text{Ł}(\check{\boldsymbol{x}}, \boldsymbol{y}_1)\} = \left[\nabla^{[1,1]} \text{Ł}(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}_1)\right] \mathsf{Cov}\{\boldsymbol{y}\} \left[\nabla^{[1,1]} \text{Ł}(\check{\boldsymbol{x}}, \bar{\boldsymbol{y}}_1)\right]'.$$

For more accurate approximations, the methods used in [57] may be useful.

If $\beta = 0$, then $\hat{\boldsymbol{x}}$ is simply the ML estimator, and then in many cases it will turn out that $\check{\boldsymbol{x}} = \boldsymbol{x}_{\text{true}}$ and (23.15.1) simplifies to the inverse of the Fisher information matrix, corresponding to the well-known asymptotic efficiency of ML estimators.

However, when $\beta \neq 0$, the estimator $\hat{\boldsymbol{x}}$ is not asymptotically efficient but (23.15.1) nevertheless describes the asymptotic covariance and $\check{\boldsymbol{x}}$ is the asymptotic mean. Furthermore, penalized-likelihood estimators are optimal with respect to the uniform CR bound for biased estimators [56].

x,cov,robust

Example 23.15.1 *The simplest such problem would be the unregularized case with a scalar unknown and i.i.d. scalar measurements:*

$$\Psi(\theta, \boldsymbol{y}) = \sum_{n=1}^{N} \psi(\theta, y_n).$$

Define

$$\check{\theta} = \arg\min_{\theta} \mathsf{E}[\psi(\theta, Y)]$$

then $\sqrt{N}(\hat{\theta} - \check{\theta})$ *is asymptotically normal with variance*

$$\mathsf{Var}\Big\{\dot{\psi}\big(\check{\theta}, Y\big)\Big\} / \mathsf{E}^2\Big[\ddot{\psi}(\theta, Y)\Big].$$

For comparison, the approximation (23.3.12) developed in [4] suggests

$$\mathsf{Var}\Big\{\hat{\theta}\Big\} \approx \frac{\mathsf{Var}\{Y\}}{N} \frac{[\nabla^{[1,1]}\, \psi\big(\check{\theta}, \bar{Y}\big)]^2}{[\ddot{\psi}\big(\check{\theta}, \bar{Y}\big)]^2}.$$

If $\psi(x, y) = \psi(x - y)$ *then* $\nabla^{[1,1]}\, \psi = \ddot{\psi}$ *in which case that approximation reduces to simply* $\mathsf{Var}\Big\{\hat{\theta}\Big\} \approx \mathsf{Var}\{Y\}/N$. *This is insufficiently accurate for nonquadratic cost functions as seen in the nonquadratic regularization results in (23.3.12).*

The following relationships may be useful for further investigations:

$$\mathsf{E}\Big[\ddot{\psi}\big(\check{\theta} - Y\big)\Big] = \int \ddot{\psi}\big(\check{\theta} - y\big)\, \mathsf{p}(y)\, \mathrm{d}y$$

$$\mathsf{Var}\Big\{\dot{\psi}\big(\check{\theta} - Y\big)\Big\} = \int [\dot{\psi}\big(\check{\theta} - y\big) - 0]^2\, \mathsf{p}(y)\, \mathrm{d}y$$

because

$$\mathsf{E}\Big[\dot{\psi}\big(\check{\theta}, Y\big)\Big] = 0.$$

s,mav,local

23.16 Local second-order statistics (s,mav,local)

23.16.1 Continuous-space random processes

Let $g(\vec{x})$ denote a continuous-space random process with mean function $\mathsf{E}[g(\vec{x})]$ and autocorrelation function

$$\mathsf{R}_g(\vec{x}, \vec{y}) = \mathsf{E}[g(\vec{x})\, g^*(\vec{y})] \,.$$

In general, such a random process need not be wide-sense stationary (*WSS*), so it need not have a power spectrum P_0. However, often for image reconstruction problems, the second-order statistics of $g(\vec{x})$ may vary slowly spatially, so it can be useful to define a *local autocorrelation function* $\mathsf{R}_0(\vec{\tau})$ and a *local power spectrum* $\mathsf{P}_0(\vec{\nu})$ near some point $\vec{x}_0$ of interest.

One intuitive definition for the *local autocorrelation function* near a point $\vec{x}_0$ is the following[6]:

$$\mathsf{R}_0(\vec{\tau}) \triangleq \mathsf{R}_g(\vec{x}_0 + \vec{\tau}, \vec{x}_0) \,.$$

An alternative definition is the following [58, p. 870]:

e,mav,local,1/2

$$\mathsf{R}_0(\vec{\tau}) \triangleq \mathsf{R}_g(\vec{x}_0 + \vec{\tau}/2, \vec{x}_0 - \vec{\tau}/2) \,. \tag{23.16.1}$$

In fact, to generalize these definitions one might consider an entire family of the form

e,mav,local,alf

$$\mathsf{R}_0(\vec{\tau}) \triangleq \mathsf{R}_g(\vec{x}_0 + \alpha\vec{\tau}, \vec{x}_0 - (1-\alpha)\vec{\tau}) \tag{23.16.2}$$

for various $\alpha \in [0, 1]$. Note that for any such definition we have that

$$\mathsf{R}_0\left(\vec{0}\right) = \mathsf{R}_g(\vec{x}_0, \vec{x}_0) = \mathsf{E}\left[|g(\vec{x}_0)|^2\right],$$

so all such definitions lead to the correct variance expression. This property is important for approximation accuracy because the *Cauchy-Schwarz inequality* ensures that autocorrelation functions are peaked at the origin $\vec{0}$.

The local WSS approximation that follows from defining a local autocorrelation function is the following:

$$\mathsf{R}_g(\vec{x}, \vec{y}) \approx \mathsf{R}_0(\vec{x} - \vec{y})$$

provided both $\vec{x}$ and $\vec{y}$ are "sufficiently close" to the point $\vec{x}_0$.

We define the *local power spectrum* as the Fourier transform of the local autocorrelation function:

$$\mathsf{P}_0(\vec{\nu}) \triangleq \int \mathsf{R}_0(\vec{\tau})\, \mathrm{e}^{-\imath 2\pi \vec{\nu}\cdot\vec{\tau}}\, \mathrm{d}\vec{\tau} \,.$$

Note that when we choose $\alpha = 1/2$, *i.e.*, (23.16.1), then $\mathsf{R}_0(\cdot)$ is symmetric, *i.e.*, $\mathsf{R}_0(-\vec{\tau}) = \mathsf{R}_0(\vec{\tau})$. This ensures that the local power spectrum approximation P_0 is real (when $g(\vec{x})$ is real). Other choices of α need not ensure this desirable property.

23.16.2 Discrete-space random processes

Now consider a $\bar{d}$-dimensional discrete-space random process $g[\vec{n}]$ for $\vec{n} \in \mathbb{Z}^{\bar{d}}$ having autocorrelation function

$$\mathsf{R}_g[\vec{n}, \vec{m}] = \mathsf{E}[g[\vec{n}]\, g[\vec{m}]] \,.$$

A generalized definition of local autocorrelation function like (23.16.2) seems inapplicable because here R_g is defined only on the integers. Therefore the natural definition of the local autocorrelation function near a point $\vec{n}_0$ is

$$\mathsf{R}_0[\vec{n}] \triangleq \mathsf{R}_g[\vec{n}_0 + \vec{n}, \vec{n}_0] \,.$$

If it happens that the form of $\mathsf{R}_g[\vec{n}, \vec{m}]$ (or an approximation thereof) can be defined sensibly for non-integer arguments (even though $g[n, m]$ itself cannot), then we can make a generalized definition of the local autocorrelation function as follows:

$$\mathsf{R}_0[\vec{n}] \triangleq \mathsf{R}_g[\vec{n}_0 + \alpha\vec{n}, \vec{n}_0 - (1-\alpha)\vec{n}]$$

for some $\alpha \in [0, 1]$. (Such a situation arises in §25.7.12 for example.) Given any such definition, there are *two* useful approximations to the local power spectrum, described next.

[6] When defining $\mathsf{R}_0(\cdot)$, our notation suppresses the dependence on the point $\vec{x}_0$. Nevertheless, this dependence is quite important.

For the first case, if $g[\vec{n}]$ is an image with finite support $N_1 \times N_2 \times \cdots \times N_d$, then we may want to consider $g[\vec{n}]$ to be a cyclicly WSS random process and interpret all of the arguments above modulo N_k. In this case the natural definition of power spectrum uses a $\vec{N}$-point discrete Fourier transform (*DFT*) as follows[7]:

$$\mathsf{P}_0\left[\vec{k}\right] = \sum_{\vec{n}=-\vec{N}/2}^{\vec{N}/2-1} \mathsf{R}_0[\vec{n}]\, \mathrm{e}^{-\imath 2\pi \vec{n}\cdot(\vec{k}/\vec{N})}\,.$$

Alternatively, we might consider $g[\vec{n}]$ to be defined over all of $\mathbb{Z}^{\bar{d}}$ (regardless of whether it really is), in which case the natural definition of power spectrum uses a discrete-space Fourier transform (*DSFT*) as follows:

$$\mathsf{P}_0(\vec{\omega}) = \sum_{\vec{n}} \mathsf{R}_0[\vec{n}]\, \mathrm{e}^{-\imath \vec{\omega}\cdot\vec{n}}\,.$$

Both definitions are merely approximations when $g[\vec{n}]$ is not WSS.

Note again that if we can choose $\alpha = 1/2$, then R_0 is symmetric and P_0 is real.

23.17 Bias of data-weighted least-squares (*DWLS*) (s,mav,dwls,bias)

s,mav,dwls,bias

As detailed in [41], *data-weighted least-squares* (*DWLS*) cost functions can lead to significant biases for Poisson data.

See [59] for alternative cost functions such as

$$\sum_{i=1}^{n_{\mathrm{d}}} (y_i + \min(y_i, 1) - \bar{y}_i)^2 / (y_i + 1).$$

[7] Above, $\vec{N} = (N_1, N_2, \ldots, N_{\bar{d}})$ and "$\vec{k}/\vec{N}$" denotes element-wise division.

s,mav,crb

23.18 *Cramér-Rao* bound (s,mav,crb)

This section presents a succinct derivation of the *Cramer-Rao bound* (*CRB*) for both unbiased and biased estimators. Covariance bounds are useful for establishing performance limits of estimators, and for imaging system design.

Consider a noisy measurement vector $\boldsymbol{y} \in \mathbb{R}^{n_\mathrm{d}}$ whose distribution $\mathsf{p}(\boldsymbol{y};\boldsymbol{x})$ depends on a parameter vector $\boldsymbol{x} \in \mathbb{R}^{n_\mathrm{p}}$. The log likelihood for $\boldsymbol{x}$ is $\log \mathsf{p}(\boldsymbol{y};\boldsymbol{x})$. We denote the $n_\mathrm{p} \times 1$ column gradient of the log-likelihood as

e,mav,crb,lgrad

$$\boldsymbol{g} = \boldsymbol{g}(\boldsymbol{x},\boldsymbol{y}) = \nabla_{\boldsymbol{x}} \log \mathsf{p}(\boldsymbol{y};\boldsymbol{x}) = \frac{1}{\mathsf{p}(\boldsymbol{y};\boldsymbol{x})} \nabla_{\boldsymbol{x}} \mathsf{p}(\boldsymbol{y};\boldsymbol{x}) \,. \tag{23.18.1}$$

Let $\hat{\boldsymbol{x}} = \hat{\boldsymbol{x}}(\boldsymbol{y})$ denote an estimator for $\boldsymbol{x}$, and let $\boldsymbol{\mu} = \boldsymbol{\mu}(\boldsymbol{x})$ denote its expectation[8]:

$$\boldsymbol{\mu}(\boldsymbol{x}) = \mathsf{E}[\hat{\boldsymbol{x}}(\boldsymbol{y})] = \int \hat{\boldsymbol{x}}(\boldsymbol{y}) \, \mathsf{p}(\boldsymbol{y};\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} \,.$$

We say that $\hat{\boldsymbol{x}}$ is an *unbiased estimator* if $\boldsymbol{\mu}(\boldsymbol{x}) = \boldsymbol{x}$. However, in image formation problems, rarely do we have unbiased estimators, so it is important to analyze the case of biased estimators.

The following theorem gives two important properties of the log-likelihood gradient $\boldsymbol{g}$.

t,mav,crb,prop

Theorem 23.18.1 *Under suitable regularity conditions, the log-likelihood gradient $\boldsymbol{g}$ in (23.18.1) satisfies:*

$$\begin{aligned} \mathsf{E}[\boldsymbol{g}] &= \boldsymbol{0} \\ \mathsf{E}[\hat{\boldsymbol{x}}\, \boldsymbol{g}'] &= \mathsf{Cov}\{\hat{\boldsymbol{x}}, \boldsymbol{g}\} = \nabla \boldsymbol{\mu}(\boldsymbol{x}). \end{aligned}$$

Proof:
Assuming $\mathsf{p}(\boldsymbol{y};\boldsymbol{x})$ is sufficiently regular to allow exchange of the order of integration and differentiation [60, p. 118]:

$$\begin{aligned} \mathsf{E}[\boldsymbol{g}] &= \int [\nabla_{\boldsymbol{x}} \log \mathsf{p}(\boldsymbol{y};\boldsymbol{x})] \, \mathsf{p}(\boldsymbol{y};\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} \\ &= \int \left[\frac{1}{\mathsf{p}(\boldsymbol{y};\boldsymbol{x})} \nabla_{\boldsymbol{x}} \mathsf{p}(\boldsymbol{y};\boldsymbol{x}) \right] \mathsf{p}(\boldsymbol{y};\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} \\ &= \int \nabla_{\boldsymbol{x}} \mathsf{p}(\boldsymbol{y};\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} = \nabla_{\boldsymbol{x}} \int \mathsf{p}(\boldsymbol{y};\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} = \nabla_{\boldsymbol{x}} 1 = \boldsymbol{0}. \end{aligned}$$

Similarly

$$\begin{aligned} \nabla \boldsymbol{\mu}(\boldsymbol{x}) &= \nabla_{\boldsymbol{x}} \int \hat{\boldsymbol{x}}(\boldsymbol{y}) \, \mathsf{p}(\boldsymbol{y};\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} = \int \hat{\boldsymbol{x}}(\boldsymbol{y}) \left[\frac{1}{\mathsf{p}(\boldsymbol{y};\boldsymbol{x})} \nabla_{\boldsymbol{x}} \mathsf{p}(\boldsymbol{y};\boldsymbol{x}) \right] \mathsf{p}(\boldsymbol{y};\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} \\ &= \int \hat{\boldsymbol{x}}(\boldsymbol{y}) \;\, \boldsymbol{g}'(\boldsymbol{x},\boldsymbol{y}) \, \mathsf{p}(\boldsymbol{y};\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} = \mathsf{E}[\hat{\boldsymbol{x}}\, \boldsymbol{g}'] = \mathsf{Cov}\{\hat{\boldsymbol{x}}, \boldsymbol{g}\}, \end{aligned}$$

where the latter equality follows because $\boldsymbol{g}$ is zero mean. □

Using Theorem 23.18.1, we can now derive the CRB. We first define the *Fisher information matrix*:

e,mav,crb,F

$$\mathbf{F} = \mathbf{F}(\boldsymbol{x}) = \mathsf{E}[\boldsymbol{g}\, \boldsymbol{g}'] = \int [\nabla \log \mathsf{p}(\boldsymbol{y};\boldsymbol{x})] \, [\nabla \log \mathsf{p}(\boldsymbol{y};\boldsymbol{x})] \, \mathsf{p}(\boldsymbol{y};\boldsymbol{x}) \, \mathrm{d}\boldsymbol{y} = \mathsf{Cov}\{\boldsymbol{g}\} \,. \tag{23.18.2}$$

The covariance of any random vector is necessarily positive-semidefinite, *i.e.*, $\mathsf{Cov}\{\boldsymbol{z}\} \succeq \boldsymbol{0}$. With the benefit of hindsight, we examine the particular vector $\hat{\boldsymbol{x}} - \boldsymbol{S}\, \boldsymbol{g}$, where $\boldsymbol{S}$ is any $n_\mathrm{p} \times n_\mathrm{p}$ matrix, as follows:

$$\boldsymbol{0} \preceq \mathsf{Cov}\{\hat{\boldsymbol{x}} - \boldsymbol{S}\, \boldsymbol{g}\} = \mathsf{Cov}\{\hat{\boldsymbol{x}}\} - \mathsf{Cov}\{\hat{\boldsymbol{x}}, \boldsymbol{g}\} \, \boldsymbol{S}' - \boldsymbol{S} \, \mathsf{Cov}\{\boldsymbol{g}, \hat{\boldsymbol{x}}\} + \boldsymbol{S} \, \mathsf{Cov}\{\boldsymbol{g}\} \, \boldsymbol{S}'.$$

Rearranging yields the following lower bound for *any* estimator $\hat{\boldsymbol{x}}$:

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \succeq \nabla \boldsymbol{\mu} \, \boldsymbol{S}' + \boldsymbol{S} \; (\nabla \boldsymbol{\mu})' - \boldsymbol{S} \mathbf{F} \boldsymbol{S}',$$

which holds for any $n_\mathrm{p} \times n_\mathrm{p}$ matrix $\boldsymbol{S}$. We get the desired lower bound by choosing $\boldsymbol{S} = \nabla \boldsymbol{\mu} \, \mathbf{F}^\dagger$, which yields the following CRB for *biased* estimators:

e,mav,crb

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \succeq \nabla \boldsymbol{\mu} \, \mathbf{F}^\dagger \; (\nabla \boldsymbol{\mu})' \,. \tag{23.18.3}$$

Another form for this bound uses the bias:

$$\boldsymbol{b}(\boldsymbol{x}) = \mathsf{E}[\hat{\boldsymbol{x}}] - \boldsymbol{x} = \boldsymbol{\mu}(\boldsymbol{x}) - \boldsymbol{x},$$

with its corresponding *bias gradient* matrix

$$\nabla \boldsymbol{b} = \nabla \boldsymbol{\mu} - \boldsymbol{I}.$$

[8] To be explicit, we could write $\mathsf{E}_{\boldsymbol{x}}[\cdot]$ throughout, to note that the expectation depends on the true value of the parameter $\boldsymbol{x}$.

Thus we can write the biased CRB as follows:

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \succeq [\boldsymbol{I} + \nabla \boldsymbol{b}]\ \mathsf{F}^{\dagger}\ [\boldsymbol{I} + \nabla \boldsymbol{b}]'.$$

If $\hat{\boldsymbol{x}}$ is an unbiased estimator for $\boldsymbol{x}$, then $\boldsymbol{\mu}(\boldsymbol{x}) = \boldsymbol{x}$ so $\nabla \boldsymbol{\mu} = \boldsymbol{I}$ and $\nabla \boldsymbol{v} = \boldsymbol{0}$, leading to the conventional CRB for unbiased estimators:

e,mav,crb,unbiased

$$\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \succeq \mathsf{F}^{\dagger}. \tag{23.18.4}$$

x,mav,crb,lin

Example 23.18.2 *Consider the linear signal model with additive gaussian noise:* $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{\varepsilon}$, *where* $\boldsymbol{\varepsilon} \sim \mathsf{N}(\boldsymbol{0}, \boldsymbol{\Pi})$, *with corresponding Fisher information matrix* $\mathsf{F} = \boldsymbol{A}'\boldsymbol{\Pi}^{-1}\boldsymbol{A}'$.

Consider any *linear estimator* $\hat{\boldsymbol{x}} = \boldsymbol{L}\boldsymbol{y}$ *for this problem. For example, for the quadratically penalized weighted least squares estimator*

$$\hat{\boldsymbol{x}} = \arg\min_{\boldsymbol{x}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{x}\|^2_{\boldsymbol{W}^{1/2}} + \boldsymbol{x}'\boldsymbol{R}\boldsymbol{x} = \left[\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A} + \boldsymbol{R}\right]^{-1} \boldsymbol{A}'\boldsymbol{W}\boldsymbol{y},$$

we would have $\boldsymbol{L} = \left[\boldsymbol{A}'\boldsymbol{W}\boldsymbol{A} + \boldsymbol{R}\right]^{-1} \boldsymbol{A}'\boldsymbol{W}$.

Then in general the covariance is $\mathsf{Cov}\{\hat{\boldsymbol{x}}\} = \boldsymbol{L}\boldsymbol{\Pi}\boldsymbol{L}'$ *and the mean gradient is* $\nabla \boldsymbol{\mu} = \boldsymbol{L}\boldsymbol{A}$. *So the biased CRB (23.18.3)* for this estimator *is*

$$\boldsymbol{L}\boldsymbol{\Pi}\boldsymbol{L}' = \mathsf{Cov}\{\hat{\boldsymbol{x}}\} \succeq \boldsymbol{L}\boldsymbol{A}\left[\boldsymbol{A}'\boldsymbol{\Pi}^{-1}\boldsymbol{A}\right]^{\dagger} \boldsymbol{A}'\boldsymbol{L}'.$$

Unfortunately, this inequality does not shed any insight on how to "best" choose the reconstruction matrix $\boldsymbol{L}$.

As a curiosity, considering the case $\boldsymbol{L} = \boldsymbol{I}$ *leads to the following interesting matrix inequality:*

$$\boldsymbol{\Pi} \succeq \boldsymbol{A}\left[\boldsymbol{A}'\boldsymbol{\Pi}^{-1}\boldsymbol{A}\right]^{\dagger} \boldsymbol{A}'.$$

Unfortunately, the biased CRB is of very limited practical use because it is *estimator dependent*; *i.e.*, it only applies to estimators with a given mean gradient $\nabla \boldsymbol{\mu}$, which is usually a very small family of estimators. So the biased CRB cannot be used as a fundamental benchmark against which to compare a wide variety of estimation methods. It could be used for imaging system design if one is willing to design the system around a given estimator.

The uniform CR bounds considered in [56], [61] overcome this limitation by bounding estimation variance subject to *constraints* on the allowable bias.

For the case of complex parameter vectors, see [62].

23.19 CRB with equality constraints

s,mav,crb,constrian

Following [51], [52], [63], if we have two sets of parameters $\boldsymbol{x} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}$ then the CRB is $\mathsf{Cov}\{\hat{\boldsymbol{x}}\} \succeq \begin{bmatrix} \mathsf{F}_{11} & \mathsf{F}_{12} \\ \mathsf{F}_{12}' & \mathsf{F}_{22} \end{bmatrix}^{-1}$.

Using the block matrix inverse formula (26.1.11): $\mathsf{Cov}\{\hat{\boldsymbol{x}}_1\} \succeq \left[\mathsf{F}_{11} - \mathsf{F}_{12}\mathsf{F}_{22}^{-1}\mathsf{F}_{12}'\right]^{-1}$. If we know $\boldsymbol{x}_2$ (*i.e.*, if we have an equality constraint for $\boldsymbol{x}_2$) then we need only estimate $\boldsymbol{x}_1$ and we get the following constrained CRB: $\mathsf{Cov}\{\hat{\boldsymbol{x}}_1\} \succeq \mathsf{F}_{11}^{-1}$. Note that $\left[\mathsf{F}_{11} - \mathsf{F}_{12}\mathsf{F}_{22}^{-1}\mathsf{F}_{12}'\right]^{-1} \succeq \mathsf{F}_{11}^{-1}$, so incorporating equality constraints reduces the covariance lower bound for unbiased estimators.

23.20 Problems (s,mav,prob)

s,mav,prob

p,mav,nec

Problem 23.1 *A widely-used metric of PET scanner performance is* noise equivalent counts *(*NEC*). Stearns analyzed FBP to relate NEC to local image noise [64]. Use the techniques in this chapter and Chapter 22 to relate NEC and spatial resolution (local impulse response) for penalized-likelihood or PWLS image reconstruction methods. (Need typed.)*

p,mav,iter

Problem 23.2 *Prove that the iteration-dependent covariance expression (23.7.1) converges to the "fixed-point" expression (23.3.14) under suitable conditions on* ϕ. *(Solve?)*

23.21 Bibliography

fessler:00:edo

[1] J. A. Fessler, H. Erdoğan, and W. B. Wu, "Exact distribution of edge-preserving MAP estimators for linear signal models with Gaussian measurement noise," *IEEE Trans. Im. Proc.*, vol. 9, no. 6, 1049–56, Jun. 2000. DOI: 10.1109/83.846247 (cit. on p. 23.2).

chesler:77:ndt

[2] D. A. Chesler, S. J. Riederer, and N. J. Pelc, "Noise due to photon statistics in computed X-ray tomography," *J. Comp. Assisted Tomo.*, vol. 1, no. 1, 64–74, Jan. 1977. [Online]. Available: http://gateway.ovid.com/ovidweb.cgi?T=JS&MODE=ovid&NEWS=n&PAGE=toc&D=ovft&AN=00004728-197701000-00009 (cit. on p. 23.2).

luenberger:69
[3] D. G. Luenberger, *Optimization by vector space methods*. New York: Wiley, 1969. [Online]. Available: http://books.google.com/books?id=lZU0CAH4RccC (cit. on p. 23.2).

fessler:96:mav
[4] J. A. Fessler, "Mean and variance of implicitly defined biased estimators (such as penalized maximum likelihood): Applications to tomography," *IEEE Trans. Im. Proc.*, vol. 5, no. 3, 493–506, Mar. 1996. DOI: 10.1109/83.491322 (cit. on pp. 23.2, 23.4, 23.5, 23.7, 23.9, 23.12, 23.22).

kim:04:ibi
[5] J. Kim and J. A. Fessler, "Intensity-based image registration using robust correlation coefficients," *IEEE Trans. Med. Imag.*, vol. 23, no. 11, 1430–44, Nov. 2004. DOI: 10.1109/TMI.2004.835313 (cit. on p. 23.2).

serfling:80
[6] R. J. Serfling, *Approximation theorems of mathematical statistics*. New York: Wiley, 1980 (cit. on pp. 23.2, 23.3, 23.13, 23.18).

hero:92:acr
[7] A. O. Hero, "A Cramer-Rao type lower bound for essentially unbiased parameter estimation," Lincoln Laboratory, MIT, Tech. Rep. 890, Jan. 1992 (cit. on p. 23.2).

fessler:93:crl
[8] J. A. Fessler and A. O. Hero, "Cramer-Rao lower bounds for biased image reconstruction," in *Proc. Midwest Symposium on Circuits and Systems*, vol. 1, 1993, 253–6. DOI: 10.1109/MWSCAS.1993.343082 (cit. on p. 23.2).

fessler:95:moi
[9] 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.*, vol. 4, 1995, 2291–4. DOI: 10.1109/ICASSP.1995.479949 (cit. on p. 23.2).

segal:94:vfm
[10] 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*, vol. 56, no. 2, 345–52, 1994. [Online]. Available: http://www.jstor.org/stable/info/2345905?seq=1 (cit. on p. 23.2).

cox:90:aao
[11] D. D. Cox and F. O'Sullivan, "Asymptotic analysis of penalized likelihood and related estimators," *Ann. Stat.*, vol. 18, no. 4, 1676–95, 1990. DOI: 10.1214/aos/1176347872 (cit. on pp. 23.2, 23.13).

osullivan:86:asp
[12] F. O'Sullivan, "A statistical perspective on ill-posed inverse problems," *Statist. Sci.*, vol. 1, no. 4, 502–27, 1986. DOI: 10.1214/ss/1177013525 (cit. on p. 23.2).

kerfoot:99:tao
[13] I. B. Kerfoot and Y. Bresler, "Theoretical analysis of an information theoretic algorithm for vector field segmentation," *IEEE Trans. Im. Proc.*, vol. 8, no. 6, 798–820, Jun. 1999. DOI: 10.1109/83.766858 (cit. on p. 23.3).

lalush:95:afa
[14] D. S. Lalush and B. M. W. Tsui, "A fast and stable maximum a posteriori conjugate gradient reconstruction algorithm," *Med. Phys.*, vol. 22, no. 8, 1273–84, Aug. 1995. DOI: 10.1118/1.597614 (cit. on p. 23.3).

qi:99:fco
[15] J. Qi and R. M. Leahy, "Fast computation of the covariance of MAP reconstructions of PET images," in *Proc. SPIE 3661 Medical Imaging 1999: Image. Proc.*, 1999, 344–55. DOI: 10.1117/12.348589 (cit. on p. 23.3).

qi:00:ran
[16] ——, "Resolution and noise properties of MAP reconstruction for fully 3D PET," *IEEE Trans. Med. Imag.*, vol. 19, no. 5, 493–506, May 2000. DOI: 10.1109/42.870259 (cit. on p. 23.3).

liang:93:cfa
[17] Z. Liang, "Compensation for attenuation, scatter, and detector response in SPECT reconstruction via iterative FBP methods," *Med. Phys.*, vol. 20, no. 4, 1097–106, Jul. 1993. DOI: 10.1118/1.597006 (cit. on p. 23.3).

hudson:94:air
[18] H. M. Hudson and R. S. Larkin, "Accelerated image reconstruction using ordered subsets of projection data," *IEEE Trans. Med. Imag.*, vol. 13, no. 4, 601–9, Dec. 1994. DOI: 10.1109/42.363108 (cit. on p. 23.3).

trussell:83:ccf
[19] H. J. Trussell, "Convergence criteria for iterative restoration methods," *IEEE Trans. Acoust. Sp. Sig. Proc.*, vol. 31, no. 1, 129–36, Feb. 1983. DOI: 10.1109/TASSP.1983.1164013 (cit. on p. 23.3).

barrett:94:npo
[20] H. H. Barrett, D. W. Wilson, and B. M. W. Tsui, "Noise properties of the EM algorithm: I. Theory," *Phys. Med. Biol.*, vol. 39, no. 5, 833–46, May 1994. DOI: 10.1088/0031-9155/39/5/004 (cit. on pp. 23.3, 23.9, 23.14).

wilson:94:npo
[21] D. W. Wilson, B. M. W. Tsui, and H. H. Barrett, "Noise properties of the EM algorithm: II. Monte Carlo simulations," *Phys. Med. Biol.*, vol. 39, no. 5, 847–72, May 1994. DOI: 10.1088/0031-9155/39/5/005 (cit. on pp. 23.3, 23.9, 23.14).

soares:00:nco
[22] E. J. Soares, C. L. Byrne, and S. J. Glick, "Noise characterization of block-iterative reconstruction algorithms. I. Theory," *IEEE Trans. Med. Imag.*, vol. 19, no. 4, 261–70, Apr. 2000. DOI: 10.1109/42.848178 (cit. on p. 23.3).

qi:03:aun
[23] J. Qi, "A unified noise analysis for iterative image estimation," *Phys. Med. Biol.*, vol. 48, no. 21, 3505–20, Nov. 2003. DOI: 10.1088/0031-9155/48/21/004 (cit. on p. 23.3).

soares:05:nco
[24] E. J. Soares, S. J. Glick, and J. W. Hoppin, "Noise characterization of block-iterative reconstruction algorithms: II. Monte Carlo simulations," *IEEE Trans. Med. Imag.*, vol. 24, no. 1, 112–21, Jan. 2005. DOI: 10.1109/TMI.2004.836876 (cit. on p. 23.3).

andrews:99:ewa
[25] D. W. K. Andrews, "Estimation when a parameter is on a boundary," *Econometrica*, vol. 67, no. 6, 1341–83, Nov. 1999. DOI: 10.1111/1468-0262.00082 (cit. on p. 23.3).

vandervaart:99 [26] A. Van der Vaart, *Asympotic statistics*. Cambridge: Cambridge, 1999 (cit. on pp. 23.3, 23.4).

huber:81 [27] P. J. Huber, *Robust statistics*. New York: Wiley, 1981 (cit. on pp. 23.3, 23.4, 23.18, 23.21).

helstrom:64:tda [28] C. W. Helstrom, "The detection and resolution of optical signals," *IEEE Trans. Info. Theory*, vol. 10, no. 4, 275–87, Oct. 1964. DOI: 10.1109/TIT.1964.1053702 (cit. on p. 23.3).

rao:73 [29] C. R. Rao, *Linear statistical inference and its applications*. New York: Wiley, 1973 (cit. on p. 23.5).

williamson:72 [30] R. E. Williamson, R. H. Crowell, and H. F. Trotter, *Calculus of vector functions*. New Jersey: Prentice-Hall, 1972 (cit. on p. 23.5).

young:71 [31] D. M. Young, *Iterative solution of large linear systems*. New York: Academic Press, 1971 (cit. on pp. 23.6, 23.14).

li:04:aeo [32] Q. Li, E. Asma, J. Qi, J. R. Bading, and R. M. Leahy, "Accurate estimation of the Fisher information matrix for the PET image reconstruction problem," *IEEE Trans. Med. Imag.*, vol. 23, no. 9, 1057–64, Sep. 2004. DOI: 10.1109/TMI.2004.833202 (cit. on p. 23.7).

fessler:96:srp [33] J. A. Fessler and W. L. Rogers, "Spatial resolution properties of penalized-likelihood image reconstruction methods: Space-invariant tomographs," *IEEE Trans. Im. Proc.*, vol. 5, no. 9, 1346–58, Sep. 1996. DOI: 10.1109/83.535846 (cit. on p. 23.7).

zhangoconnor:07:fpo [34] Y. Zhang-O'Connor and J. A. Fessler, "Fast predictions of variance images for fan-beam transmission tomography with quadratic regularization," *IEEE Trans. Med. Imag.*, vol. 26, no. 3, 335–46, Mar. 2007. DOI: 10.1109/TMI.2006.887368 (cit. on p. 23.8).

zhangoconnor:07:fvp [35] ——, "Fast variance predictions for 3D cone-beam CT with quadratic regularization," in *Proc. SPIE 6510 Medical Imaging 2007: Phys. Med. Im.*, 2007, 65105W:1–10. DOI: 10.1117/12.710312 (cit. on p. 23.8).

valenzuela:10:phd [36] J. Valenzuela, "Polarimetric image reconstruction algorithms," PhD thesis, Univ. of Michigan, Ann Arbor, MI, 48109-2122, Ann Arbor, MI, 2010 (cit. on p. 23.9).

li:11:npf [37] Y. Li, "Noise propagation for iterative penalized-likelihood image reconstruction based on Fisher information," *Phys. Med. Biol.*, vol. 56, no. 4, 1083–104, Feb. 2011. DOI: 10.1088/0031-9155/56/4/013 (cit. on p. 23.9).

carson:94:paa [38] 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 Trans. Med. Imag.*, vol. 13, no. 3, 526–37, Sep. 1994. DOI: 10.1109/42.310884 (cit. on pp. 23.9, 23.14).

lange:84:era [39] K. Lange and R. Carson, "EM reconstruction algorithms for emission and transmission tomography," *J. Comp. Assisted Tomo.*, vol. 8, no. 2, 306–16, Apr. 1984 (cit. on pp. 23.13, 23.15).

sauer:93:alu [40] K. Sauer and C. Bouman, "A local update strategy for iterative reconstruction from projections," *IEEE Trans. Sig. Proc.*, vol. 41, no. 2, 534–48, Feb. 1993. DOI: 10.1109/78.193196 (cit. on p. 23.13).

fessler:95:hpp [41] J. A. Fessler, "Hybrid Poisson/polynomial objective functions for tomographic image reconstruction from transmission scans," *IEEE Trans. Im. Proc.*, vol. 4, no. 10, 1439–50, Oct. 1995. DOI: 10.1109/83.465108 (cit. on pp. 23.13, 23.14, 23.17, 23.18, 23.24).

bouman:93:fnm [42] C. Bouman and K. Sauer, "Fast numerical methods for emission and transmission tomographic reconstruction," in *Proc. 27th Conf. Info. Sci. Sys., Johns Hopkins*, 1993, 611–6 (cit. on pp. 23.14, 23.17).

bouman:96:aua [43] C. A. Bouman and K. Sauer, "A unified approach to statistical tomography using coordinate descent optimization," *IEEE Trans. Im. Proc.*, vol. 5, no. 3, 480–92, Mar. 1996. DOI: 10.1109/83.491321 (cit. on pp. 23.14, 23.17).

clinthorne:93:pmf [44] 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 Trans. Med. Imag.*, vol. 12, no. 1, 78–83, Mar. 1993. DOI: 10.1109/42.222670 (cit. on p. 23.14).

huesman:84:anf [45] R. H. Huesman, "A new fast algorithm for the evaluation of regions of interest and statistical uncertainty in computed tomography," *Phys. Med. Biol.*, vol. 29, no. 5, 543–52, May 1984. DOI: 10.1088/0031-9155/29/5/007 (cit. on pp. 23.15, 23.18).

fessler:95:pml [46] J. A. Fessler and A. O. Hero, "Penalized maximum-likelihood image reconstruction using space-alternating generalized EM algorithms," *IEEE Trans. Im. Proc.*, vol. 4, no. 10, 1417–29, Oct. 1995. DOI: 10.1109/83.465106 (cit. on pp. 23.15, 23.16).

lange:90:coe [47] K. Lange, "Convergence of EM image reconstruction algorithms with Gibbs smoothing," *IEEE Trans. Med. Imag.*, vol. 9, no. 4, 439–46, Dec. 1990, Corrections, T-MI, 10:2(288), June 1991. DOI: 10.1109/42.61759 (cit. on p. 23.16).

kak:88 [48] A. C. Kak and M. Slaney, *Principles of computerized tomographic imaging*. New York: IEEE Press, 1988. DOI: 10.1137/1.9780898719277. [Online]. Available: http://www.slaney.org/pct (cit. on p. 23.17).

fessler:94:pwl
[49] J. A. Fessler, "Penalized weighted least-squares image reconstruction for positron emission tomography," *IEEE Trans. Med. Imag.*, vol. 13, no. 2, 290–300, Jun. 1994. DOI: 10.1109/42.293921 (cit. on p. 23.17).

lalush:93:afa
[50] 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)*, vol. 34, no. 5, p. 27, May 1993 (cit. on p. 23.17).

gorman:90:lbf
[51] J. D. Gorman and A. O. Hero, "Lower bounds for parametric estimators with constraints," *IEEE Trans. Info. Theory*, vol. 36, no. 6, 1285–301, Nov. 1990. DOI: 10.1109/18.59929 (cit. on pp. 23.18, 23.26).

marzetta:93:asd
[52] T. L. Marzetta, "A simple derivation of the constrained multiple parameter Cramer-Rao bound," *IEEE Trans. Sig. Proc.*, vol. 41, no. 6, 2247–9, Jun. 1993. DOI: 10.1109/78.218151 (cit. on pp. 23.18, 23.26).

carson:93:aaf
[53] 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 Trans. Med. Imag.*, vol. 12, no. 2, 240–50, Jun. 1993. DOI: 10.1109/42.232252 (cit. on p. 23.18).

fessler:95:rpo
[54] J. A. Fessler, "Resolution properties of regularized image reconstruction methods," Comm. and Sign. Proc. Lab., Dept. of EECS, Univ. of Michigan, Ann Arbor, MI, 48109-2122, Tech. Rep. 297, Aug. 1995. [Online]. Available: http://web.eecs.umich.edu/~fessler/papers/lists/files/tr/95,297,rpo.pdf (cit. on p. 23.18).

snyder:93:irf
[55] D. L. Snyder, A. M. Hammoud, and R. L. White, "Image recovery from data acquired with a charge-coupled-device camera," *J. Opt. Soc. Am. A*, vol. 10, no. 5, 1014–23, May 1993. DOI: 10.1364/JOSAA.10.001014 (cit. on p. 23.18).

eldar:04:mvi
[56] Y. C. Eldar, "Minimum variance in biased estimation: Bounds and asymptotically optimal estimators," *IEEE Trans. Sig. Proc.*, vol. 52, no. 7, 1915–30, Jul. 2004. DOI: 10.1109/TSP.2004.828929 (cit. on pp. 23.21, 23.26).

field:82:ssa
[57] C. A. Field, "Small sample asymptotic expansions for multivariate M-estimates," *Ann. Stat.*, vol. 10, no. 3, 672–89, Sep. 1982. DOI: 10.1214/aos/1176345864 (cit. on p. 23.21).

barrett:03
[58] H. H. Barrett and K. J. Myers, *Foundations of image science*. New York: Wiley, 2003 (cit. on p. 23.23).

mighell:99:pei
[59] K. Mighell, "Parameter estimation in astronomy with Poisson-distributed data. I. The chi-square-lambda statistic," *The Astrophysical Journal*, vol. 518, 380–93, Jun. 1999. DOI: 10.1086/307253 (cit. on p. 23.24).

lehmann:98
[60] E. L. Lehmann and G. Casella, *Theory of point estimation*. New York: Springer-Verlag, 1998 (cit. on p. 23.25).

hero:96:eeb
[61] A. O. Hero, J. A. Fessler, and M. Usman, "Exploring estimator bias-variance tradeoffs using the uniform CR bound," *IEEE Trans. Sig. Proc.*, vol. 44, no. 8, 2026–41, Aug. 1996. DOI: 10.1109/78.533723 (cit. on p. 23.26).

abeida:05:gcr
[62] H. Abeida and J.-P. Delmas, "Gaussian Cramer-Rao bound for direction estimation of noncircular signals in unknown noise fields," *IEEE Trans. Sig. Proc.*, vol. 53, no. 12, 4610–8, Dec. 2005. DOI: 10.1109/TSP.2005.859226 (cit. on p. 23.26).

stoica:98:otc
[63] P. Stoica and B. C. Ng, "On the Cramer-Rao bound under parametric constraints," *IEEE Signal Proc. Letters*, vol. 5, no. 7, 177–9, Jul. 1998. DOI: 10.1109/97.700921 (cit. on p. 23.26).

stearns:04:nal
[64] C. W. Stearns, "NEC and local image noise in PET imaging," in *Proc. IEEE Nuc. Sci. Symp. Med. Im. Conf.*, 2004, M5–278. DOI: 10.1109/NSSMIC.2004.1466338 (cit. on p. 23.26).