1. Uncertainty

1.1. **Uncertainty in ML and MAP estimates.** You will often hear me say that reporting how well you believe an estimate is just as important as reporting an estimate (You will hear other say it too, and once you get bitten by believing an estimate you had no business believing, you'll probably start saying it too). There are at least two concepts of uncertainty that can be easy to confuse unless you carefully think about it. The first concept of uncertainty is the "per measurement uncertainty". For example, the covariance of the conditional probability density function is an example of what I'm calling this "intra-measurement uncertainty." It tells you that given this particular measured data (and implicity your model assumptions) what can you really say about what you are trying to estimate. It helps you form confidence regions and interval estimates. The second concept of uncertainty is the "inter-measurement uncertainty." This tells you how much you expect your conditional probability density function (and in particular it's mean, peak, and/or covariance) to change across different measurements. For example, suppose you've selected the mean of the conditional probability density function as your estimate $\tilde{\mathbf{f}}_{MMSE}(\mathbf{g})$ and consider the covariance of the conditional probability funciton $\mathbf{C}_{\mathbf{f}|\mathbf{g}}(\mathbf{g})$ as the "per-measurement uncertainty." These are both dependent on the data $\mathbf{g}$ and therefore will change as the data changes.

If reporting a single estimated parameter, we need to know something about how much change we can expect in this estimate as different data sets are acquired. In fact, this is often how uncertainty is dealt with in many estimation applications where (Monte Carlo) simulated data is constructed based on a model of the object and the measurement process and then the estimator is applied. This process is repeated several times until a satisfactory knowledge of the statistics of the the estimator is obtained (generally mean and covariance). These are the emipirical means and covariances of the estimator. If the estimator is $\tilde{\mathbf{f}}$, then the most-often sought statistics are the mean $E\tilde{\mathbf{f}}$ and covariance $\mathbf{C}_{\tilde{\mathbf{f}}} = E\left(\tilde{\mathbf{f}} - E\tilde{\mathbf{f}}\right)\left(\tilde{\mathbf{f}} - E\tilde{\mathbf{f}}\right)^T$ (quite often only the diagonal elements of this matrix—the variances on the individual elements of $\tilde{\mathbf{f}}$—are of interest). The expectaction is "taken" over $\mathbf{g}$ in these expressions (*i.e.* the expectaction integral is $p_{\mathbf{g}}(\mathbf{g})$).

Quite often we are concerned with the total error covariance,

$$\mathbf{C}_e = E\left(\tilde{\mathbf{f}} - \mathbf{f}\right)\left(\tilde{\mathbf{f}} - \mathbf{f}\right)^T,$$

The diagonal terms of this matrix are the mean-squared error terms for individual components of $\mathbf{f}$. The trace of this matrix is the (total) mean-squared error

$$e^2 = \text{trace}\mathbf{C}_e = E\left(\tilde{\mathbf{f}} - \mathbf{f}\right)^T\left(\tilde{\mathbf{f}} - \mathbf{f}\right)$$

It can be shown that

$$\mathbf{C}_e = \mathbf{C}_{\tilde{\mathbf{f}}} + \mathbf{b}\mathbf{b}^T$$

where $\mathbf{b} = E\tilde{\mathbf{f}} - E\mathbf{f}$ is the bias. So, the error covariance involves both the bias and the covariance of the estimate. This relationship is often written just for the diagonal elements of the preceeding matrix equation

$$\text{MSE}\left(\tilde{f}_i\right) = \text{var}\left(\tilde{f}_i\right) + \text{bias}^2\left(\tilde{f}_i\right)$$

which states that the mean-square error is the combination of the variance of the estimator (the precision) and the bias (the accuracy). We will look at this uncertainty for the Bayesian estimator first and then consider the maximum-likelihood estimator

1.2. **Matrix inequalities.** In order to understand error on random vectors it is important to get some familiarity with what it means for two covariance matrices to be compared.

1.2.1. *Bayesian.* Consider first the unimodal, symmetrical conditional PDF case (*e.g.* Gaussian) where the Bayesian estimate and non-linear MMSE estimate coincide. Then, $\tilde{\mathbf{f}}_{MMSE} = E\{\mathbf{f}|\mathbf{g}\}$. Using the fact that

$$E\mathbf{f} = E\{E\{\mathbf{f}|\mathbf{g}\}\}$$

we immediately conclude that this estimate is unbiased (note, however, that for the common case of assuming zero-mean statistics for the prior on $\mathbf{f}$ this says that averaging over all realizations of the data the estimator will return an average of zero) and therefore the trace of the covariance matrix is also minimized by the MMSE estimate. What about the covariance matrix of the MMSE estimate?

$$\mathbf{C}_{\mathbf{f}|\mathbf{g}} = E\left\{\left(\mathbf{f} - \tilde{\mathbf{f}}_{\text{MMSE}}\right)\left(\mathbf{f} - \tilde{\mathbf{f}}_{\text{MMSE}}\right)^T |\mathbf{g}\right\}$$

and using iterated expectations we see that

$$\begin{aligned} \mathbf{C}_e &= E\left\{E\left\{\left(\mathbf{f} - \tilde{\mathbf{f}}_{\text{MMSE}}\right)\left(\mathbf{f} - \tilde{\mathbf{f}}_{\text{MMSE}}\right)^T |\mathbf{g}\right\}\right\} \\ &= E\mathbf{C}_{\mathbf{f}|\mathbf{g}}. \end{aligned}$$

So, the covariance matrix of the estimator (averaged over $\mathbf{g}$) is the same as the error covariance of the estimator. This is a nice result that connects the two notions of uncertainty. One can think of the covariance matrix as providing a single-sample estimate of the covariance of the error.

We saw how to get $\mathbf{C}_{\mathbf{f}|\mathbf{g}}$ for Gaussian statistics. What can we do in the more general case? One strategy that works whenever it makes sense to report a single $\mathbf{f}$ anyway is to approximate the covariance of the estimate by assuming that, locally at least, the conditional probability density function curve can be modeled as Gaussian. In other words in a neighborhood around the peak $\tilde{\mathbf{f}}$, we expand the conditional distribution function in a second-order Tayler-series. Thus,

$$\begin{aligned}\log p_{\mathbf{f}|\mathbf{g}}(\mathbf{f}|\mathbf{g}) &\approx \log p_{\mathbf{f}|\mathbf{g}}\left(\tilde{\mathbf{f}}|\mathbf{g}\right) + \left.\frac{\partial \log p_{\mathbf{f}|\mathbf{g}}}{\partial f_i}\right|_{\tilde{\mathbf{f}}}\left(f_i - \hat{f}_i\right) + \frac{1}{2}\left.\frac{\partial \log p_{\mathbf{f}|\mathbf{g}}}{\partial f_i \partial f_j}\right|_{\tilde{\mathbf{f}}}\left(f_i - \tilde{f}_i\right)\left(f_j - \tilde{f}_j\right) \\ &\approx \log p_{\mathbf{f}|\mathbf{g}}\left(\tilde{\mathbf{f}}|\mathbf{g}\right) + \frac{1}{2}\left.\frac{\partial \log p_{\mathbf{f}|\mathbf{g}}}{\partial f_i \partial f_j}\right|_{\tilde{\mathbf{f}}}\left(f_i - \tilde{f}_i\right)\left(f_j - \tilde{f}_j\right).\end{aligned}$$

The first-order term was eliminated by recognizing that at the maximum, the gradient of the function being maximized goes to zero. Identifying these terms with the equivalent terms if $p_{\mathbf{f}|\mathbf{g}}$ represented a Gaussian distribution, we see that

$$-\left.\frac{\partial \log p_{\mathbf{f}|\mathbf{g}}}{\partial f_i \partial f_j}\right|_{\tilde{\mathbf{f}}_M}$$

plays the role of an approximate inverse covariance matrix (it is the exact inverse covariance matrix if $p_{\mathbf{f}|\mathbf{g}}$ is actually Gaussian). In terms of the likelihood function, the approximate covariance matrix is

$$\left(\mathbf{C}_{\mathbf{f}|\mathbf{g}}^{-1}\right)_{ij} \approx -\frac{\partial \log p_{\mathbf{g}|\mathbf{f}}}{\partial f_i \partial f_j} - \frac{\partial \log p_{\mathbf{f}}}{\partial f_i \partial f_j},$$

which shows explicitly how the assumed information about $\mathbf{f}$ goes to alter the covariance in the estimate.

1.2.2. *Maximum-Likelihood estimates.* Recall that the limiting form of maximizing the conditional PDF of $\mathbf{f}|\mathbf{g}$ for the case of no-information in $\mathbf{f}$ (infinite variance) gives the maximum-likelihood (ML) estimates. In this limit, because the derivatives of $p_{\mathbf{f}}$ tend to zero, we can use the Taylor expansion of the previous section to approximate an uncertainty for maximum-likelihood estimates.

$$\left(\mathbf{C}_{\mathbf{f}|\mathbf{g}}^{-1}\right)_{ij} \approx -\frac{\partial \log p_{\mathbf{f}|\mathbf{g}}}{\partial f_i \partial f_j} = -\frac{\partial \log p_{\mathbf{g}|\mathbf{f}}}{\partial f_i \partial f_j} = -\frac{\partial L_{\mathbf{g}}(\mathbf{f})}{\partial f_i \partial f_j}$$

where all of the derivatives are evaluated at $\hat{\mathbf{f}}_{ML}$. Notice that for inverse-problems, this inverse covariance matrix is typically not full rank or poorly conditioned (eigenvalues near zero) so that it's inverse (the covariance matrix of $\mathbf{f}|\mathbf{g}$) shows just how bad our ML estimate is. This is corrected with the Bayes estimator by adding the term

$$\frac{\partial \log p_{\mathbf{f}}}{\partial f_i \partial f_j}$$

to the inverse covariance (provided that near the maximum there is enough change in this second-derivative matrix)

For understanding the bias of ML estimates $\mathbf{f}$ is understood be a degenerate random-variable centered at $\mathbf{f}_0$ with zero covariance (a delta-function PDF). Then $E\mathbf{f} = \mathbf{f}_0$ and $\mathbf{b} = E\tilde{\mathbf{f}}_{ML} - \mathbf{f}_0$. It is known that ML estimates are asymptotically unbiased in that as the number of independent measurements of $\mathbf{g}$ increases, the bias tends to 0.

For inverse problems ML estimates are generally not useful except as a means to develop iterative algorithms which we then truncate early. The main reason we have spent time with them in this section is because the expected value of the approximate conditional probability density function shown above plays an important role. Recall that for the case of the MMSE estimator we showed that

$$E\mathbf{C}_{\mathbf{f}|\mathbf{g}} = \mathbf{C}_e.$$

For the approximation to $\mathbf{C}_{\mathbf{f}|\mathbf{g}}$ derived in this section, the Cramer-Rao bound shows that

$$x_i E\left(\mathbf{C}_{\mathbf{f}|\mathbf{g}}^{-1}\right)_{ij} x_j \approx x_i E\left\{-\frac{\partial L_{\mathbf{g}}(\mathbf{f})}{\partial f_i \partial f_j}\right\} x_j \geq x_i \left(\mathbf{C}_e^{-1}\right)_{ij} x_j.$$

where $\mathbf{C}_e$ is the error-covariance matrix of **any** unbiased estimator of $\mathbf{f}$. In other words

$$\mathbf{J}_{ML}^{-1} \leq \mathbf{C}_e.$$

where

$$J_{ML,ij} = E\left\{-\frac{\partial L_{\mathbf{g}}(\mathbf{f})}{\partial f_i \partial f_j}\right\} = E\left[\frac{\partial L_{\mathbf{g}}(\mathbf{f})}{\partial f_i}\frac{\partial L_{\mathbf{g}}(\mathbf{f})}{\partial f_j}\right].$$

is called the Fisher information matrix. The equality is a classic result of estimation theory. There is a sometimes noted extension to biased estimators which states that

$$\mathbf{C}_e \geq \mathbf{b}\mathbf{b}^T + \left[\mathbf{I} + \frac{\partial \mathbf{b}}{\partial \mathbf{f}}\right]^T \mathbf{J}^{-1} \left[\mathbf{I} + \frac{\partial \mathbf{b}}{\partial \mathbf{f}}\right]$$

if $\mathbf{b}(\mathbf{f}) = E\tilde{\mathbf{f}} - \mathbf{f}$ is the bias of the estimator. In this theory we assume that $\mathbf{f}$ is a deterministic quantity and so all expectations are done over $\mathbf{g}$ alone.

1.3. **Bayesian bounds.** The biggest problem with the Cramer-Rao (CR) bound is that it does not tell us what the effect of any prior information is on our bound. We have seen that adding some (often just a little) prior information to the problem can allow us to turn ill-posed problems into well-posed ones using Bayesian techniques. Effectively we have turned a problem which had infinite uncertainty according to the Cramer-Rao bound into one with reasonable uncertainty. This has it's largest influence when we have many variables that we are estimating at a time, and the prior information we provide is the correlation (the expected statistical connection) between the variables. But, the Cramer-Rao bound has no way to help us predict the effect of this information. This is because the CR bound depends only on the likelihood function (i.e. no prior information is used). Another problem with the CR bound is that the biased version is usually very hard to apply so that in practice it is rarely useful for biased estimators.

Fortunately, other bounds are available which can incorporate prior knowledge. The Weiss-Weinstein bound (WWB) and the Ziv-Zakai bound (ZZB) are two examples. We will discuss only the WWB. The Weiss-Weinstein bound is one of a family of bounds derived using the Cauchy-Swartz, or (more generally) the Holder inequality. A Bayesian Cramer-Rao bound is also a member of this family.

1.3.1. *Minimum Mean Square Error.* We already know, theoretically, the best bound on the covariance matrix of an estimate because we have proven that $\mathbf{f}_{\mathrm{MMSE}}$ always gives the lowest covariance matrix. In other words we know that for any estimator of $\mathbf{f}$: $\tilde{\mathbf{f}} = h(\mathbf{g})$, the error covariance matrix

$$E\left\{[\mathbf{f} - h(\mathbf{g})][\mathbf{f} - h(\mathbf{g})]^T\right\} \geq E\left\{[\mathbf{f} - \mathbf{f}_{\mathrm{MMSE}}][\mathbf{f} - \mathbf{f}_{\mathrm{MMSE}}]^T\right\}$$

and so the right-hand side is a natural bound on the error covariance matrix of any estimator for $\mathbf{f}$. However, $\mathbf{f}_{\mathrm{MMSE}}$ can be very hard to determine.

When $p(\mathbf{f}|\mathbf{g})$ is Gaussian (Gaussian noise, Gaussian prior and linear model) then we know that

$$\begin{aligned}
\mathbf{f}_{\mathrm{MMSE}} &= \mathbf{C}_{\mathbf{f}|\mathbf{g}}\mathbf{A}^H\mathbf{C}^{-1}(\mathbf{g} - \mathbf{A}\mu_{\mathbf{f}}) + \mu_{\mathbf{f}} \\
\mathbf{C}_{\mathbf{f}|\mathbf{g}} &= \left(\mathbf{A}^H\mathbf{C}^{-1}\mathbf{A} + \mathbf{C}_{\mathbf{f}}^{-1}\right)^{-1}
\end{aligned}$$

and the bound on the error-covariance for any estimator is

$$\mathbf{C}_e \geq E\mathbf{C}_{\mathbf{f}|\mathbf{g}} = \left(\mathbf{A}^H\mathbf{C}^{-1}\mathbf{A} + \mathbf{C}_{\mathbf{f}}^{-1}\right)^{-1},$$

assuming that $\mathbf{C}$ and $\mathbf{C}_{\mathbf{f}}$ are known and do not depend on $\mathbf{f}$.

For comparision, under the same circumstances the Cramer-Rao bound states that any unbiased estimator not using prior information is bound by

$$\mathbf{C}_e \geq \left(\mathbf{A}^H\mathbf{C}^{-1}\mathbf{A}\right)^{-1}.$$

This shows that the effect of the prior has been to add $\mathbf{C}_{\mathbf{f}}^{-1}$ to the Fisher information matrix. For inverse problems this addition is essential to allowing the problem to be solved.

1.3.2. *Weiss-Weinstein bound.* When the noise is not Gaussian, or the prior is not Gaussian or (most commonly) the model is not linear, then it is generally not so easy to come up with the MMSE bound. Other bounds have therefore been developed. One that is not too difficult to understand is the general class of bounds discovered by Weinstein and Weiss. Using our notation, we let $\mathbf{f}$ be the vector of N_f parameters to be estimated and define $\mathbf{g}$ as the data. The joint probability density function of $\mathbf{f}$ and $\mathbf{g}$ is $p(\mathbf{g}, \mathbf{f}) = p(\mathbf{g}|\mathbf{f})\, p(\mathbf{f})$.

Let $\psi_i(\mathbf{g},\mathbf{f})$ (for $i = 1 \ldots N_f$) be functions such that for (almost) all $\mathbf{g}$:

$$\int d\mathbf{f}\psi_i(\mathbf{g},\mathbf{f})\,p(\mathbf{g},\mathbf{f}) = 0.$$

Define the matrices $\mathbf{V}$ and $\mathbf{P}$ with elements

$$\begin{aligned} V_{ij} &= E\left[f_i\psi_j(\mathbf{g},\mathbf{f})\right] \\ P_{ij} &= E\left[\psi_i(\mathbf{g},\mathbf{f})\,\psi_i(\mathbf{g},\mathbf{f})\right], \end{aligned}$$

where expectations are taken over the joint probability density function $p(\mathbf{g},\mathbf{f})$. Then, for any estimator $\tilde{\mathbf{f}} = h(\mathbf{g})$ it can be shown that

$$E\left[\left(\mathbf{f}-\tilde{\mathbf{f}}\right)\left(\mathbf{f}-\tilde{\mathbf{f}}\right)^T\right] \geq \mathbf{V}\mathbf{P}^{-1}\mathbf{V}.$$

Notice that this bound is valid for any set of functions ψ_i that we can pick. If we choose

$$\psi_i = f_i - E\left\{f_i|\mathbf{g}\right\}$$

then the WWB is a restatement of the known fact that the mean of the conditional density function is the estimator with minimum mean-squared error. Suppose instead we choose

$$\psi_i = \frac{\partial \log p(\mathbf{g},\mathbf{f})}{\partial f_i}$$

as long $p(\mathbf{g},\mathbf{f}) \neq 0$ (and $\psi_i = 0$ otherwise). We need to show that the required orthogonality condition holds

$$\begin{aligned} \int d\mathbf{f}\frac{\partial \log p(\mathbf{g},\mathbf{f})}{\partial f_i}p(\mathbf{g},\mathbf{f}) &= \int d\mathbf{f}\frac{1}{p(\mathbf{g},\mathbf{f})}\frac{\partial p(\mathbf{g},\mathbf{f})}{\partial f_i}p(\mathbf{g},\mathbf{f}) \\ &= \int d\mathbf{f}\frac{\partial p(\mathbf{g},\mathbf{f})}{\partial f_i} \\ &= 0 \end{aligned}$$

where the last result assumes that $p(\mathbf{g},\mathbf{f}) = 0$ on the boundary of the integration region in each dimension of $\mathbf{f}$.

Then,

$$
\begin{aligned}
V_{ij} &= E\left[f_i \frac{\partial \log p\left(\mathbf{g},\mathbf{f}\right)}{\partial f_j}\right] \\
&= \int d\mathbf{f} \int d\mathbf{g} f_i \frac{\partial p\left(\mathbf{g},\mathbf{f}\right)}{\partial f_j} \\
&= \int d\mathbf{f} f_i \frac{\partial p\left(\mathbf{f}\right)}{\partial f_j} \\
&= \int d\mathbf{f} \frac{\partial\left(f_i p\left(\mathbf{f}\right)\right)}{\partial f_j} - \int d\mathbf{f} \delta_{ij} p\left(\mathbf{f}\right) \\
&= \delta_{ij}
\end{aligned}
$$

assuming that $\lim_{f_i \to \pm\infty} f_i p\left(\mathbf{f}\right) = 0$ for each i. Define

$$
J_{ij} = E\left[\frac{\partial \log p\left(\mathbf{g},\mathbf{f}\right)}{\partial f_i} \frac{\partial \log p\left(\mathbf{g},\mathbf{f}\right)}{\partial f_j}\right],
$$

then this version of the Weiss-Weinstein bound suggests that

$$
\mathbf{C}_e \geq \mathbf{J}^{-1}.
$$

This could be called the Bayesian information matrix. Notice there are no restrictions about the estimator being biased or unbiased for this bound. We can relate this bound to our previous discussion by noting that

$$
\begin{aligned}
J_{ij} &= E\left[\frac{\partial \log p\left(\mathbf{g},\mathbf{f}\right)}{\partial f_i} \frac{\partial \log p\left(\mathbf{g},\mathbf{f}\right)}{\partial f_j}\right] = -E\left[\frac{\partial^2 \log p\left(\mathbf{g},\mathbf{f}\right)}{\partial f_i \partial f_j}\right] \\
&= -E\left[\frac{\partial^2 \log p\left(\mathbf{g}|\mathbf{f}\right)}{\partial f_i \partial f_j} + \frac{\partial^2 \log p\left(\mathbf{f}\right)}{\partial f_i \partial f_j}\right] \\
&= E\left[\mathbf{J}_{ML}\right] + \mathbf{J}_p
\end{aligned}
$$

where $\mathbf{J}_{ML}$ is the Fisher-information matrix (calculated with expectations taken over $\mathbf{g}$ with $\mathbf{f}$ fixed – i.e. $\mathbf{g}|\mathbf{f}$) and

$$
\mathbf{J}_p = -E\left[\frac{\partial^2 \log p\left(\mathbf{f}\right)}{\partial f_i \partial f_j}\right].
$$

Broken up in this fashion, it shows that the Bayesian information matrix alters the Fisher information matrix in two ways. First, it averages the Fisher information matrix over all $\mathbf{f}$ and second, it shows how the prior increases the information in the estimate.

1.4. **Uncertainty for an estimate that is the maximum (or minimum) of any objective function.** If our estimate $\tilde{\mathbf{f}}$ is formed by maximizing some objective function (notice this includes the ML and Bayesian cases)

$$
\tilde{\mathbf{f}} = \arg\max_{\mathbf{f}} \Phi_{\mathbf{g}}\left(\mathbf{f}\right),
$$

where $\mathbf{g}$ is the data, then we can treat this as if it were a Bayesian estimation problem for the purposes of understanding the "intra-measurement" unceratinty in $\tilde{\mathbf{f}}$. In other-words we recognize that this approach is equivalent to maximizing the conditional PDF of $p_{\mathbf{f}|\mathbf{g}} = C \exp\left[\Phi_{\mathbf{g}}\left(\mathbf{f}\right)\right]$ where C is an appropriate normalizing constant. Therefore, we can expand the log of $p_{\mathbf{f}|\mathbf{g}}$ locally in a second-order Taylor-series about the peak to find that the equivalent "local" covariance is approximately

$$\left(\mathbf{C}_{\mathbf{f}|\mathbf{g}}^{-1}\right)_{ij} \approx -\left.\frac{\partial \Phi_{\mathbf{g}}\left(\mathbf{f}\right)}{\partial f_i \partial f_j}\right|_{\tilde{\mathbf{f}}}.$$

Notice that this approximate error can be computed completely from the data.

What about inter-measurement uncertainty? Can we approximate the covariance matrix for general M-estimates (estimates which are extreme-values of some function). There is an approach introduced by Fessler which can be useful (although pure statisticians might balk).

The idea is to recognize that finding the maximum of a function is the same as solving the equation

$$\left.\frac{\partial \Phi_{\mathbf{g}}\left(\mathbf{f}\right)}{\partial \mathbf{f}}\right|_{\tilde{\mathbf{f}}} = 0$$

(where we've assumed that $\mathbf{f}$ is real-valued for now). This is an implict equation generally non-linear equation in terms of the data $\mathbf{g}$

$$\tilde{\mathbf{f}} = \mathbf{h}\left(\mathbf{g}\right).$$

Assuming that the noise on the data is not too high (read: this can break down badly in low SNR cases), then expand this function in a Taylor series around the mean of the data

$$\tilde{\mathbf{f}} = \mathbf{h}\left(\bar{\mathbf{g}}\right) + \left.\frac{\partial \mathbf{h}}{\partial g_i}\right|_{\bar{\mathbf{g}}}\left(g_i - \bar{g}_i\right) + \frac{\partial^2 \mathbf{h}}{\partial g_i \partial g_j}\left(g_i - \bar{g}_i\right)\left(g_j - \bar{g}_j\right).$$

The mean of $\tilde{\mathbf{f}}$ is then

$$\mathbf{h}\left(\bar{\mathbf{g}}\right) + \frac{\partial^2 \mathbf{h}}{\partial g_i \partial g_j}\left(\mathbf{C}_{\mathbf{g}}\right)_{ij}$$

(often the second term can be dropped), while the covariance of $\tilde{\mathbf{f}}$ is to-first order

$$\begin{aligned}\mathbf{C}_{\tilde{\mathbf{f}}} &= \frac{\partial \mathbf{h}}{\partial g_i}\left(\mathbf{C}_{\mathbf{g}}\right)_{ij}\frac{\partial \mathbf{h}}{\partial g_j}^T \\ &= \mathbf{D}\mathbf{C}_{\mathbf{g}}\mathbf{D}^T\end{aligned}$$

The derivatives (which should be evaluated at $\tilde{\mathbf{f}}$) can be calculated using implicit differentiation. We have

$$\frac{\partial \Phi_{\mathbf{g}}\left(\mathbf{h}\left(\mathbf{g}\right)\right)}{\partial f_n} = 0$$

Thus taking the derivative of this equation with respect to g_i we get

$$\begin{aligned} 0 &= \frac{\partial}{\partial g_i}\left[\frac{\partial \Phi_{\mathbf{g}}\left(\mathbf{h}\left(\mathbf{g}\right)\right)}{\partial f_n}\right] \\ &= \frac{\partial^2 \Phi_{\mathbf{g}}\left(\tilde{\mathbf{f}}\right)}{\partial g_i \partial f_n} + \frac{\partial^2 \Phi_{\mathbf{g}}\left(\tilde{\mathbf{f}}\right)}{\partial f_i \partial f_n}\frac{\partial \mathbf{h}}{\partial g_i}. \end{aligned}$$

Writing this as

$$D_{ni}^{11} + D_{ni}^{20} D_i^1$$

we see that

$$\mathbf{D}^1 = \left\{\frac{\partial \mathbf{h}}{\partial g_i}\right\} = -\left(\mathbf{D}^{20}\right)^{-1}\mathbf{D}^{11}$$

and

$$\mathbf{C}_{\tilde{\mathbf{f}}} = \mathbf{D}^{11}\left(\mathbf{D}^{20}\right)^{-1}\mathbf{C}_{\mathbf{g}}\left(\mathbf{D}^{20}\right)^{-T}\left(\mathbf{D}^{11}\right)^{T}.$$

2. Examples

An example may be useful to clarify some of the concepts.

Consider the case where $\mathbf{f}^T = [f_1, f_2]$ is two numbers and we observe

$$g = f_1^2 + f_2^2 + \nu$$

where ν is zero-mean Gaussian random-variable with variance σ_ν^2. One interpretation of this model is that we've measured the amplitude-squared of some complex number and are trying to estimate the complex number. Of course, in this case we can immediately see that this problem is not well-posed as there is not a unique solution. In more complicated cases it may not be immediately obvious. Suppose, we take a maximum likelihood approach and seek to find the $\mathbf{f}$ that maximizes, $L_g(\mathbf{f})$ for a given measurement

$$\tilde{\mathbf{f}}_{ML} = \arg\max_{\mathbf{f}}\left(-\frac{1}{2\sigma_\nu^2}\left(g - \mathbf{f}^T\mathbf{f}\right)^2\right).$$

Clearly, this is maximum whenever $f_1^2 + f_2^2 = g$, and so as expected, the ML estimator does nothing to get rid of the ill-posedness of the operator. What does the estimate of covariance give

$$\left(\mathbf{C}_{\mathbf{f}|g}^{-1}\right)_{ij} \approx -\left.\frac{\partial \Phi_g\left(\mathbf{f}\right)}{\partial f_i \partial f_j}\right|_{\tilde{\mathbf{f}}} = -\frac{2\delta_{ij}}{\sigma_\nu^2}\left[g - \left(\tilde{f}_1^2 + \tilde{f}_2^2\right)\right] + \left.\frac{4}{\sigma_\nu^2}\tilde{f}_i\tilde{f}_j\right|_{\tilde{\mathbf{f}}}$$

but $\tilde{f}_1^2 + \tilde{f}_2^2 = g$. Thus,

$$\mathbf{C}_{\mathbf{f}|g}^{-1} \approx \frac{4\tilde{f}_1\tilde{f}_2}{\sigma_\nu^2}\begin{bmatrix} \frac{\tilde{f}_1}{\tilde{f}_2} & 1 \\ 1 & \frac{\tilde{f}_2}{\tilde{f}_1} \end{bmatrix}$$

which is not an invertible matrix (one of the eigenvalues is zero) implying that at least one axis of the confidence ellipsoid is infinite (the essential characteristic of an inverse problem).

Clearly more information is needed to get a useful answer. What if, for example, we knew before-hand that $\mathbf{f}$ was positive and "near" the point (m_1, m_2). This nearness could be captured by stating that the prior PDF of $\mathbf{f}$ is Gaussian with mean $\mathbf{m} = [m_1, m_2]^T$ and covariance

$$\mathbf{C_f} = \sigma_f^2 \mathbf{I}.$$

Using a Bayes estimator we compute the estimate as

$$\tilde{\mathbf{f}}_B = \arg\max_{\mathbf{f}} \left\{ \left(-\frac{1}{2\sigma^2} \left(g - \mathbf{f}^T \mathbf{f} \right)^2 \right) - \frac{1}{2\sigma_f^2} \left(\mathbf{f} - \mathbf{m} \right)^T \left(\mathbf{f} - \mathbf{m} \right) \right\}.$$

The solution to this problem is given by

$$\tilde{f}_1 = \tilde{f}_2 \frac{m_1}{m_2}$$

with $\tilde{f}_2$ satisfying the polynomial

$$\tilde{f}_2^3 - \frac{(g - r/2)}{1 + m_1^1/m_2^2} \tilde{f}_2 - \frac{m_2 r/2}{1 + m_1^2/m_2^2} = 0,$$

and

$$r = \frac{\sigma_\nu^2}{\sigma_f^2}.$$

The approximate conditional covariance matrix is

$$\left(\mathbf{C}_{\mathbf{f}|\mathbf{g}}^{-1} \right)_{ij} \approx -\frac{\partial \Phi_{\mathbf{g}} (\mathbf{f})}{\partial f_i \partial f_j} = \frac{1}{\sigma_\nu^2} \left[4 f_i f_j + \frac{m_2 r}{\tilde{f}_2} \delta_{ij} \right].$$

In particular

$$\mathbf{C}_{\mathbf{f}|\mathbf{g}}^{-1} = 4 \frac{\tilde{f}_1 \tilde{f}_2}{\sigma_\nu^2} \begin{bmatrix} \frac{m_1}{m_2} + \frac{m_2 r}{4 \tilde{f}_2^2 f_1} & 1 \\ 1 & \frac{m_2}{m_1} + \frac{m_1 r}{4 \tilde{f}_1^2 \tilde{f}_2} \end{bmatrix}.$$