Least Squares with a Data Kernel Involving an Unknown Parameter
Bill Menke, November 19, 2020

Part 1. The idea is to view a standard "linear" inverse problem of the form $\mathbf{d}^{obs} = \mathbf{G}\mathbf{m}$ has having a data kernel $\mathbf{G}$ that is dependent on a parameter p, and then to solve for $\mathbf{m}$ and p in a way that makes use of the fact that for fixed p the solution for $\mathbf{m}$ is $\mathbf{m}(p) = [\mathbf{G}^{\mathrm{T}}\mathbf{G}]^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{d}^{obs}$.

The prediction error $\mathbf{e}(p)$ is a function of a parameter p via:

$$\mathbf{e}(p) = \mathbf{d}^{obs} - \mathbf{G}(p)\,\mathbf{m}(p)$$

Here $\mathbf{d}^{obs}$ is a vector of N observed data, $\mathbf{m}$ is a vector of M model parameters, and $\mathbf{G}$ is an $\boldsymbol{N \times M}$ matrix. For fixed p, the least squares solution is that one that minimizes the total $E(\mathbf{m}) = \|\mathbf{e}\|_2^2$ and is given by:

$$\mathbf{m}(p) = \mathbf{G}^{-g}\mathbf{d}^{obs} \quad \text{with} \quad \mathbf{G}^{-g} \equiv [\mathbf{Z}(p)]^{-1}[\mathbf{G}(p)]^{\mathrm{T}} \quad \text{and} \quad \mathbf{Z}(p) = [\mathbf{G}(p)]^{\mathrm{T}}\mathbf{G}(p)$$

Here $\mathbf{G}^{-g}$ is a generalized inverse. The least squared solution $\mathbf{m}(p)$ defines a parametric curve in the space of $\mathbf{m}$. The estimated solution $(p^{est}, \mathbf{m}^{est})$ is the point of minimum error along this curve; that is $p^{est} = \mathrm{argmin}_p\, E(p)$ and $\mathbf{m}^{est} = \mathbf{m}^{est}(p^{est})$. This point can be found using Newton's method, once a procedure for calculating the derivative $\partial\mathbf{e}/\partial p$ has been established.

The derivative of the solution with respect to the parameter p is:

$$\frac{\partial \mathbf{m}}{\partial p} = \frac{\partial \mathbf{G}^{-g}}{\partial p}\mathbf{d}^{obs} \quad \text{with}$$

$$\frac{\partial \mathbf{G}^{-g}}{\partial p} = \mathbf{Z}^{-1}\left(\frac{\partial \mathbf{G}}{\partial p}\right)^{\mathrm{T}} - \mathbf{Z}^{-1}\frac{\partial \mathbf{Z}}{\partial p}\mathbf{Z}^{-1}\mathbf{G}^{\mathrm{T}} \quad \text{and} \quad \frac{\partial \mathbf{Z}}{\partial p} = \left(\frac{\partial \mathbf{G}}{\partial p}\right)^{\mathrm{T}}\mathbf{G} + \mathbf{G}^{\mathrm{T}}\frac{\partial \mathbf{G}}{\partial p}$$

Combining these equations leads to:

$$\frac{\partial \mathbf{m}}{\partial p} = \mathbf{Z}^{-1}\left(\left(\frac{\partial \mathbf{G}}{\partial p}\right)^{\mathrm{T}}\mathbf{d}^{obs} - \frac{\partial \mathbf{Z}}{\partial p}\mathbf{m}\right) = \mathbf{Z}^{-1}\left(\left(\frac{\partial \mathbf{G}}{\partial p}\right)^{\mathrm{T}}\mathbf{e} - \mathbf{G}^{\mathrm{T}}\frac{\partial \mathbf{G}}{\partial p}\mathbf{m}\right)$$

Consequently, for fixed p the solution $\mathbf{m}(p)$ and its derivative $\partial\mathbf{m}^{est}/\partial p$ satisfy linear equations involving the same $M \times M$ matrix $\mathbf{Z}$:

$$\mathbf{Z}\frac{\partial \mathbf{m}}{\partial p} = \left(\left(\frac{\partial \mathbf{G}}{\partial p}\right)^{\mathrm{T}}\left(\mathbf{d}^{obs} - \mathbf{G}\mathbf{m}\right) - \mathbf{G}^{\mathrm{T}}\frac{\partial \mathbf{G}}{\partial p}\mathbf{m}\right) \quad \text{and} \quad \mathbf{Z}\,\mathbf{m} = \mathbf{G}^{\mathrm{T}}\mathbf{d}^{obs}$$

For fixed p, the derivative of the predicted data $\mathbf{d}^{pre}$ and the error $\mathbf{e} = \mathbf{d}^{obs} - \mathbf{d}^{pre}$ are:

$$\frac{\partial \mathbf{d}^{pre}}{\partial p} = \frac{\partial \mathbf{G}}{\partial p}\mathbf{m} + \mathbf{G}\frac{\partial \mathbf{m}}{\partial p} \quad \text{and} \quad \frac{\partial \mathbf{e}}{\partial p} = \frac{\partial}{\partial p}\left(\mathbf{d}^{obs} - \mathbf{d}^{pre}\right) = -\frac{\partial \mathbf{d}^{pre}}{\partial p}$$

Two useful second derivatives are:

$$\frac{\partial^2 \mathbf{m}}{\partial p^2} = -\mathbf{Z}^{-1}\frac{\partial \mathbf{Z}}{\partial p}\mathbf{Z}^{-1}\left(\left(\frac{\partial \mathbf{G}}{\partial p}\right)^{\mathrm{T}}\mathbf{e} - \mathbf{G}^{\mathrm{T}}\frac{\partial \mathbf{G}}{\partial p}\mathbf{m}\right) +$$

$$\mathbf{Z}^{-1}\left(\left(\frac{\partial^2 \mathbf{G}}{\partial p^2}\right)^{\mathrm{T}}\mathbf{e} + \left(\frac{\partial \mathbf{G}}{\partial p}\right)^{\mathrm{T}}\frac{\partial \mathbf{e}}{\partial p} - \left(\frac{\partial \mathbf{G}}{\partial p}\right)^{\mathrm{T}}\frac{\partial \mathbf{G}}{\partial p}\mathbf{m} - \mathbf{G}^{\mathrm{T}}\frac{\partial^2 \mathbf{G}}{\partial p^2}\mathbf{m} - \mathbf{G}^{\mathrm{T}}\frac{\partial \mathbf{G}}{\partial p}\frac{\partial \mathbf{m}}{\partial p}\right)$$

$$\frac{\partial^2 E}{\partial p^2} = -2\left(\frac{\partial \mathbf{e}}{\partial p}\right)^{\mathrm{T}}\frac{\partial \mathbf{G}}{\partial p}\mathbf{m} - 4\mathbf{e}^{\mathbf{T}}\frac{\partial \mathbf{G}}{\partial p}\frac{\partial \mathbf{m}}{\partial p} - 2\left(\frac{\partial \mathbf{e}}{\partial p}\right)^{\mathrm{T}}\mathbf{G}\frac{\partial \mathbf{m}}{\partial p} - 2\mathbf{e}^{\mathbf{T}}\frac{\partial^2 \mathbf{G}}{\partial p^2}\mathbf{m} - 2\mathbf{e}^{\mathbf{T}}\mathbf{G}\frac{\partial^2 \mathbf{m}}{\partial p^2}$$

All these formulas have been verified numerically.

Newton's method can be used to iteratively improve an estimate of the solution, starting with an initial estimate $\left(p_0, \mathbf{m}(p_0)\right)$. At the nth iteration, the solution is p_n and $\mathbf{m}_n \equiv [\mathbf{Z}(p_n)]^{-1}[\mathbf{G}(p_n)]^{\mathrm{T}}\mathbf{d}^{obs}$. We now define:

$$\mathbf{F} \equiv \left.\frac{\partial \mathbf{e}}{\partial p}\right|_{\mathbf{m}_0, p_0} \quad \text{and} \quad \mathbf{f} \equiv -\mathbf{e}(\mathbf{m}_0, p_0)$$

The least square solution for $\Delta p \equiv p_{n+1} - p_n$ is then $\Delta p = [\mathbf{F}^{\mathrm{T}}\mathbf{F}]^{-1}\mathbf{F}^{\mathrm{T}}\mathbf{f}$ and the estimated parameter is $p^{est} = \lim_{n\to\infty} p_{n+1}$. This procedure can be trivially extended to the case of several parameters by adding columns to $\mathbf{F}$. If there are K such parameters, then $(K+1)$ linear equations, each the same $M \times M$ matrix $\mathbf{Z}$, must be solved at each iteration of Newton's method.

The covariance of the estimate can be approximated using a linearized approximation, $\mathbf{C}_{m,p} \approx \sigma_d^2[\mathbf{W}^{\mathrm{T}}\mathbf{W}]^{-1}$, where:

$$\mathbf{W} = \left[\mathbf{G}(p_n) \quad \left.\frac{\partial \mathbf{G}}{\partial p}\right|_{p_n}\right]$$

Here σ_d^2 is the variance of the data.

Example 1: We consider a simple curve fitting case with $M = 1$ model parameter $m \equiv m_1$ and data kernel case $G_{ij}(p) = m\, x_i^p$, where x_i is an auxiliary variable. The derivative is $\partial G_{ij}/\partial p = x_i^p \ln x_i$. In the example $(p^{true}, m^{true}) = (2.0, 1.5)$ and $N = 101$ synthetic data are uniformly spaced on the interval $(0,1)$ with uncorrelated Normally-distributed noise with uniform variance $\sigma_d^2 = (0.05)^2$ (Figure 1, red circles). The error surface $E(p,m)$ (Figure 2, colors) has a global minimum at (p^{true}, m^{true}). The parametric curve $m(p)$ (Figure 2, blue curve) passes through the global minimum. The Newton's method, begun at the point $p_0 = 0.5$, rapidly converges to the global minimum, following the parametric curve as it does so (Figure 2, green curve with triangle at each iteration). This trajectory is different than the one followed by Newton's method when both p and m are allowed to freely vary (Figure 2, yellow curve with circles at each iteration). Both trajectories rapidly converge (in about three iteration) to the global minimum. The predicted data (Figure 1, black curve) fit the data well.

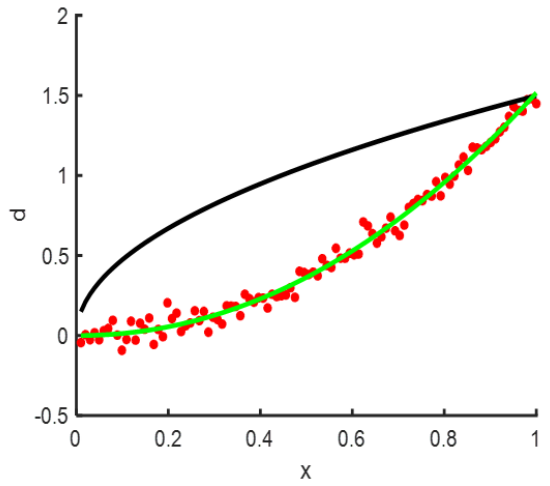

Fig. 1. First exemplary inverse problem. The observed data d (red dots) as a function of the auxiliary variable x, together with the predicted data for p_0 (black curve) and p^{est} (green curve) See text for further discussion.

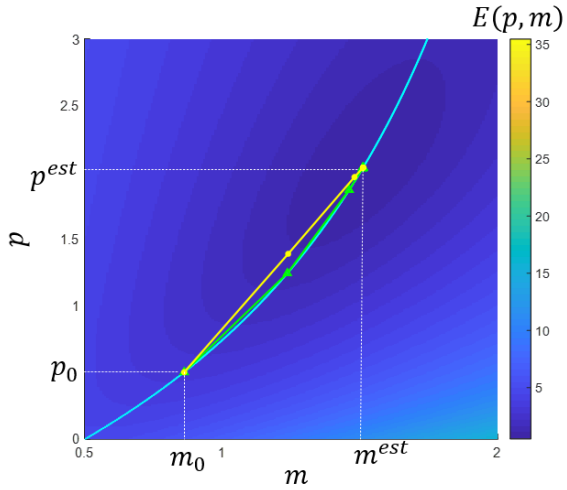

Fig. 2. Error surface (colors) for the first exemplary inverse problem in Figure 1. See text for further discussion

Example 2. In a second example, we consider a simple Fourier analysis problem, with data kernel $G_{ij}(p) = m_1 + m_2 \sin(px_i) + m_3 \cos(px_i)$ with position x on the interval $(0, 100)$, $\mathbf{m}^{true} = [1.0, 0.2, 03]^{\mathrm{T}}$ and $p^{true} = 6\pi/100$. Normally-distributed noise with variance $\sigma_d^2 = (0.05)^2$ is used to create synthetic observed data $\mathbf{d}^{obs}$ (Figure 3, red dots). The initial solution, with $p_0 = 0.9\, p^{true}$, fits the data poorly (Figure 3, black curve), whereas the solution with $p^{est} \approx 1.004\, p^{true}$ (Figure 3, green curve) fits it well.

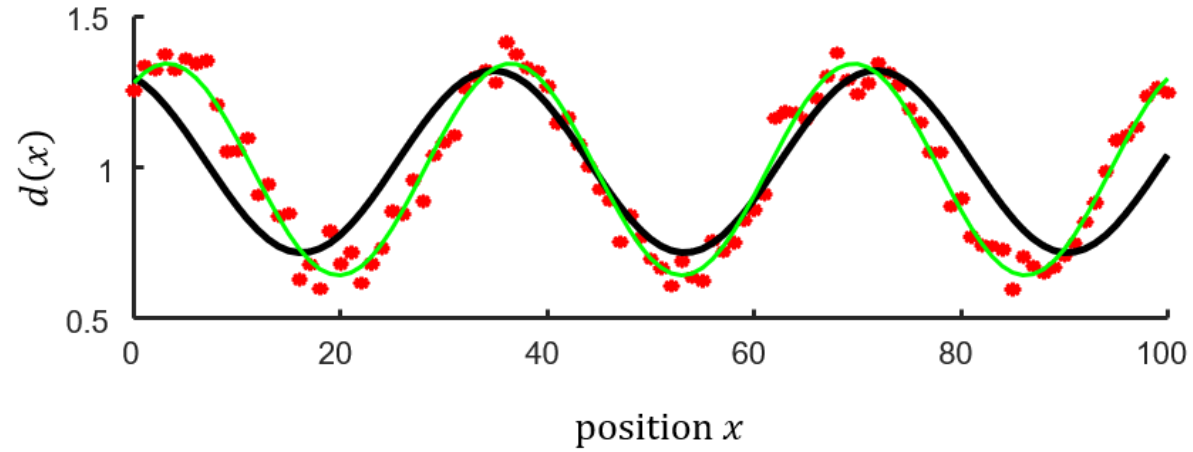

Fig. 3. Second exemplary inverse problem. The observed data d (red dots) as a function of the auxiliary variable x, together with the predicted data for p_0 (black curve) and p^{est} (green curve) See text for further discussion.

Part 2. We now derive comparable derivatives for the Generalized Least Squares solution (Menke 2018, equation 5.48):

$$\mathbf{m}^{est} = \mathbf{G}^{-g}\mathbf{d}^{obs} + \mathbf{H}^{-g}\mathbf{h}^{pri} \equiv \mathbf{m}^{(1)} + \mathbf{m}^{(2)} \quad with$$

$$\mathbf{G}^{-g} = \mathbf{Z}^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{C}_d^{-1} \text{ and } \mathbf{H}^{-g} = \mathbf{Z}^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{C}_h^{-1} \quad \text{and} \quad \mathbf{Z} = \mathbf{G}^{\mathrm{T}}\mathbf{C}_d^{-1}\mathbf{G} + \mathbf{H}^{\mathrm{T}}\mathbf{C}_h^{-1}\mathbf{H}$$

Case 1. When the data kernel $\mathbf{G}(p)$ depends on a parameter p, the derivative of the estimated model parameters is:

$$\frac{\partial \mathbf{m}^{est}}{\partial p} = \frac{\partial \mathbf{G}^{-g}}{\partial p}\mathbf{d}^{obs} + \frac{\partial \mathbf{H}^{-g}}{\partial p}\mathbf{h}^{pri}$$

$$= \frac{\partial \mathbf{Z}^{-1}}{\partial p}\mathbf{G}^{\mathrm{T}}\mathbf{C}_d^{-1}\mathbf{d}^{obs} + \mathbf{Z}^{-1}\frac{\partial \mathbf{G}^{\mathrm{T}}}{\partial p}\mathbf{C}_d^{-1}\mathbf{d}^{obs} + \frac{\partial \mathbf{Z}^{-1}}{\partial p}\mathbf{H}^{\mathrm{T}}\mathbf{C}_h^{-1}\mathbf{h}^{pri} =$$

$$= -\mathbf{Z}^{-1}\frac{\partial \mathbf{Z}}{\partial p}\mathbf{Z}^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{C}_d^{-1}\mathbf{d}^{obs} + \mathbf{Z}^{-1}\frac{\partial \mathbf{G}^{\mathrm{T}}}{\partial p}\mathbf{C}_d^{-1}\mathbf{d}^{obs} - \mathbf{Z}^{-1}\frac{\partial \mathbf{Z}}{\partial p}\mathbf{Z}^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{C}_h^{-1}\mathbf{h}^{pri} =$$

$$= \mathbf{Z}^{-1}\left\{-\frac{\partial \mathbf{Z}}{\partial p}\mathbf{m}^{(1)} + \left(\frac{\partial \mathbf{G}}{\partial p}\right)^{\mathrm{T}}\mathbf{C}_d^{-1}\mathbf{d}^{obs} - \frac{\partial \mathbf{Z}}{\partial p}\mathbf{m}^{(2)}\right\} =$$

$$= \mathbf{Z}^{-1}\left\{\left(\frac{\partial \mathbf{G}}{\partial p}\right)^{\mathrm{T}}\mathbf{C}_d^{-1}\mathbf{d}^{obs} - \frac{\partial \mathbf{Z}}{\partial p}\mathbf{m}^{(est)}\right\} =$$

$$\text{with} \quad \frac{\partial \mathbf{Z}}{\partial p} = \left(\frac{\partial \mathbf{G}}{\partial p}\right)^{\mathrm{T}}\mathbf{C}_d^{-1}\mathbf{G} + \mathbf{G}^{\mathrm{T}}\mathbf{C}_d^{-1}\frac{\partial \mathbf{G}}{\partial p}$$

The normalized prediction error is $\tilde{\mathbf{e}} = \mathbf{C}_d^{-½}\left(\mathbf{d}^{obs} - \mathbf{d}^{pre}\right) = \mathbf{C}_d^{-½}\left(\mathbf{d}^{obs} - \mathbf{G}\mathbf{m}^{est}\right)$ and its derivative is:

$$\frac{\partial \tilde{\mathbf{e}}}{\partial p} = -\mathbf{C}_d^{-½}\mathbf{G}\frac{\partial \mathbf{m}^{est}}{\partial p} - \mathbf{C}_d^{-½}\frac{\partial \mathbf{G}}{\partial p}\mathbf{m}^{est}$$

The normalized error in prior information is $\tilde{\boldsymbol{\ell}} = \mathbf{C}_h^{-½}\left(\mathbf{h}^{pri} - \mathbf{h}^{pre}\right) = \mathbf{C}_h^{-½}\left(\mathbf{h}^{pri} - \mathbf{H}\mathbf{m}^{est}\right)$ Its derivative is:

$$\frac{\partial \tilde{\boldsymbol{\ell}}}{\partial p} = -\mathbf{C}_h^{-½}\mathbf{H}\frac{\partial \mathbf{m}^{est}}{\partial p}$$

Case 2: When the prior information kernel $\mathbf{H}(p)$ depends on a parameter p, the derivatives are (by analogy):

$$\frac{\partial \mathbf{m}^{est}}{\partial p} = \frac{\partial \mathbf{G}^{-g}}{\partial p}\mathbf{d}^{obs} + \frac{\partial \mathbf{H}^{-g}}{\partial p}\mathbf{h}^{pri}$$

$$= \mathbf{Z}^{-1}\left\{\left(\frac{\partial \mathbf{H}}{\partial p}\right)^{\mathrm{T}}\mathbf{C}_h^{-1}\mathbf{h}^{pri} - \frac{\partial \mathbf{Z}}{\partial p}\mathbf{m}^{(est)}\right\} =$$

$$\text{with} \quad \frac{\partial \mathbf{Z}}{\partial p} = \left(\frac{\partial \mathbf{H}}{\partial p}\right)^{\mathrm{T}}\mathbf{C}_h^{-1}\mathbf{H} + \mathbf{H}^{\mathrm{T}}\mathbf{C}_h^{-1}\frac{\partial \mathbf{H}}{\partial p}$$

$$\frac{\partial \tilde{\mathbf{e}}}{\partial p} = -\mathbf{C}_d^{-½}\mathbf{G}\frac{\partial \mathbf{m}^{est}}{\partial p}$$

$$\frac{\partial \tilde{\boldsymbol{\ell}}}{\partial p} = -\mathbf{C}_h^{-½}\mathbf{H}\frac{\partial \mathbf{m}^{est}}{\partial p} - \mathbf{C}_h^{-½}\frac{\partial \mathbf{H}}{\partial p}\mathbf{m}^{est}$$

Case 3: When the data variance $\mathbf{C}_d(p)$ depends on a parameter p, the derivative is:

$$\frac{\partial \mathbf{m}^{est}}{\partial p} = \frac{\partial \mathbf{G}^{-g}}{\partial p}\mathbf{d}^{obs} + \frac{\partial \mathbf{H}^{-g}}{\partial p}\mathbf{h}^{pri}$$

$$= \frac{\partial \mathbf{Z}^{-1}}{\partial p}\mathbf{G}^{\mathrm{T}}\mathbf{C}_d^{-1}\mathbf{d}^{obs} + \mathbf{Z}^{-1}\mathbf{G}^{\mathrm{T}}\frac{\partial \mathbf{C}_d^{-1}}{\partial p}\mathbf{d}^{obs} + \frac{\partial \mathbf{Z}^{-1}}{\partial p}\mathbf{H}^{\mathrm{T}}\mathbf{C}_h^{-1}\mathbf{h}^{pri} =$$

$$= -\mathbf{Z}^{-1}\frac{\partial \mathbf{Z}}{\partial p}\mathbf{Z}^{-1}\mathbf{G}^{\mathrm{T}}\mathbf{C}_d^{-1}\mathbf{d}^{obs} + \mathbf{Z}^{-1}\mathbf{G}^{\mathrm{T}}\frac{\partial \mathbf{C}_d^{-1}}{\partial p}\mathbf{d}^{obs} - \mathbf{Z}^{-1}\frac{\partial \mathbf{Z}}{\partial p}\mathbf{Z}^{-1}\mathbf{H}^{\mathrm{T}}\mathbf{C}_h^{-1}\mathbf{h}^{pri} =$$

$$= -\mathbf{Z}^{-1}\frac{\partial \mathbf{Z}}{\partial p}\mathbf{m}^{(1)} + \mathbf{Z}^{-1}\mathbf{G}^{\mathrm{T}}\frac{\partial \mathbf{C}_d^{-1}}{\partial p}\mathbf{d}^{obs} - \mathbf{Z}^{-1}\frac{\partial \mathbf{Z}}{\partial p}\mathbf{m}^{(2)} =$$

$$= \mathbf{Z}^{-1}\left(\mathbf{G}^{\mathrm{T}}\frac{\partial \mathbf{C}_d^{-1}}{\partial p}\mathbf{d}^{obs} - \frac{\partial \mathbf{Z}}{\partial p}\mathbf{m}^{est}\right)$$

$$\text{with} \quad \frac{\partial \mathbf{Z}}{\partial p} = \mathbf{G}^{\mathrm{T}}\frac{\partial \mathbf{C}_d^{-1}}{\partial p}\mathbf{G} = -\mathbf{G}^{\mathrm{T}}\mathbf{C}_d^{-1}\frac{\partial \mathbf{C}_d}{\partial p}\mathbf{C}_d^{-1}\mathbf{G}$$

The derivative of the normalized prediction error is $\tilde{\mathbf{e}}$ is:

$$\frac{\partial \tilde{\mathbf{e}}}{\partial p} = -\mathbf{C}_d^{-½}\mathbf{G}\frac{\partial \mathbf{m}^{est}}{\partial p} - \frac{\partial \mathbf{C}_d^{-½}}{\partial p}\mathbf{G}\mathbf{m}^{est}$$

The derivation $\partial \mathbf{C}_d^{-½}/\partial p$ cab be computed by solving the Sylvester equation that arises from differentiating $\mathbf{C}_d^{-½}\,\mathbf{C}_d^{-½} = \mathbf{C}_d^{-1}$:

$$\frac{\partial \mathbf{C}_d^{-½}}{\partial p}\mathbf{C}_d^{-½} + \mathbf{C}_d^{-½}\frac{\partial \mathbf{C}_d^{-½}}{\partial p} = \frac{\partial \mathbf{C}_d^{-1}}{\partial p} = -\mathbf{C}_d^{-1}\frac{\partial \mathbf{C}_d}{\partial p}\mathbf{C}_d^{-1}$$

The derivative of the normalized error in prior information $\tilde{\ell}$ is:

$$\frac{\partial \tilde{\boldsymbol{\ell}}}{\partial p} = -\mathbf{C}_h^{-½}\mathbf{H}\frac{\partial \mathbf{m}^{est}}{\partial p}$$

Case 4: When the variance of prior information $\mathbf{C}_h(p)$ depends on a parameter p, the derivatives are (by analogy):

$$\frac{\partial \mathbf{m}^{est}}{\partial p} = \mathbf{Z}^{-1}\left(\mathbf{H}^{\mathrm{T}}\frac{\partial \mathbf{C}_h^{-1}}{\partial p}\mathbf{h}^{pri} - \frac{\partial \mathbf{Z}}{\partial p}\mathbf{m}^{est}\right)$$

$$\text{with} \quad \frac{\partial \mathbf{Z}}{\partial p} = \mathbf{H}^{\mathrm{T}} \frac{\partial \mathbf{C}_h^{-1}}{\partial p} \mathbf{H} = -\mathbf{H}^{\mathrm{T}} \mathbf{C}_h^{-1} \frac{\partial \mathbf{C}_h}{\partial p} \mathbf{C}_h^{-1} \mathbf{H}$$

$$\frac{\partial \tilde{\mathbf{e}}}{\partial p} = -\mathbf{C}_d^{-½} \mathbf{G} \frac{\partial \mathbf{m}^{est}}{\partial p}$$

$$\frac{\partial \tilde{\boldsymbol{\ell}}}{\partial p} = -\mathbf{C}_h^{-½} \mathbf{H} \frac{\partial \mathbf{m}^{est}}{\partial p} - \frac{\partial \mathbf{C}_h^{-½}}{\partial p} \mathbf{H} \mathbf{m}^{est}$$

$$\frac{\partial \mathbf{C}_h^{-½}}{\partial p} \mathbf{C}_h^{-½} + \mathbf{C}_h^{-½} \frac{\partial \mathbf{C}_h^{-½}}{\partial p} = -\mathbf{C}_h^{-1} \frac{\partial \mathbf{C}_h}{\partial p} \mathbf{C}_h^{-1}$$

I have not yet performed a numerical verification of these formula.