Gaussian Process Regression = Generalized Least Squares

Bill Menke, February 25, 2020
(inspired by a conversation with Roger Creel)

Part 1. Relationship between Gaussian Process Estimation and Generalized Least Squares.

Generalized Least Squares (GLS) has two equivalent formulas. One looks like ordinary least squares solution, in the sense that it contains the matrix sequence $[\mathbf{G}^T\mathbf{G}]^{-1}\mathbf{G}^T$, where $\mathbf{G}$ is the data kernel (Menke 2018, eqn. 5.39). The other looks like the minimum length solution, in the sense that it contains the matrix sequence $\mathbf{G}^T[\mathbf{G}\mathbf{G}^T]^{-1}$ (Menke, 2018, eqn. 5.38). Here, I demonstrate that the Gaussian Process Regression (GPR) method (Rasmussen et al., 2016) is a special case of the second formula of Generalized Least Squares method. The equivalence is demonstrated by making the choice $\mathbf{G} = \mathbf{I}$ and by dividing the model parameters $\mathbf{m}$ into two groups, a target group for which there is no associated data, and a training group with associated data $\mathbf{d}$.

The GLS equation is (Menke, 2018, eqn. 5.38):

$$\mathbf{m}^{\text{est}} = \langle \mathbf{m} \rangle + \mathbf{G}^{-g}\left(\mathbf{d}^{\text{obs}} - \mathbf{G}\langle \mathbf{m} \rangle\right) = \mathbf{G}^{-g}\mathbf{d}^{\text{obs}} + [\mathbf{I} - \mathbf{R}]\langle \mathbf{m} \rangle$$

$$\text{with} \quad \mathbf{G}^{-g} = [\text{cov } \mathbf{m}]_A \mathbf{G}^T \left\{[\text{cov } \mathbf{d}] + [\text{cov } \mathbf{g}] + \mathbf{G}[\text{cov } \mathbf{m}]_A \mathbf{G}^T\right\}^{-1}$$

Making the following substitutions:

$$\mathbf{m}^{\text{est}} = \bar{\mathbf{f}} \text{ and } \langle \mathbf{m} \rangle = \boldsymbol{\mu} \text{ and } \mathbf{G} = \mathbf{I} \text{ and } [\text{cov } \mathbf{m}]_A = \mathbf{K} \text{ and } [\text{cov } \mathbf{d}] = \sigma^2\mathbf{I}$$

$$\text{and } [\text{cov } \mathbf{g}] = 0 \text{ and } \mathbf{d}^{\text{obs}} = \mathbf{y}$$

leads to:

$$\mathbf{G}^{-g} = \mathbf{K}\{\mathbf{K} + \sigma^2\mathbf{I}\}^{-1}$$

$$\bar{\mathbf{f}} = \boldsymbol{\mu} + \mathbf{G}^{-g}(\mathbf{y} - \boldsymbol{\mu}) = \boldsymbol{\mu} + \mathbf{K}\{\mathbf{K} + \sigma^2\mathbf{I}\}^{-1}(\mathbf{y} - \boldsymbol{\mu})$$

Comparing this result to the one for Gaussian Process Regression (GPR) (Rasmussen et al., 2016):

$$\bar{f}^* = \mu^* + K(X^*, X)[K(X, X) + \sigma_n^2 I]^{-1}(y - \mu)$$

we find that the two are identical, except that the GPR formula distinguishes model parameters that have no associated data (the target, indicated with a star) from those that are supported by observations (the training data, indicated by the lack of a star). This difference can be eliminated by considering the GLS model parameters ordered into a first stared group with no associated data and a second primed group with associated data. We define a rectangular data kernel $\mathbf{G}$:

$$\mathbf{G} = [\mathbf{0} \quad \mathbf{I}]$$

so that $\mathbf{G}\mathbf{m} = \mathbf{d}'$, a complementary rectangular selection matrix $\mathbf{M}$ as:

$$\mathbf{M} = [\mathbf{I} \quad \mathbf{0}]$$

so that $\mathbf{M}\mathbf{m} = \mathbf{m}^*$. We also define a covariance matrix:

$$[\mathrm{cov}\,\mathbf{m}]_{\mathrm{A}} = \begin{bmatrix} \mathbf{K}(\mathrm{X}^*,\mathrm{X}^*) & \mathbf{K}(\mathrm{X}^*,\mathrm{X}') \\ \mathbf{K}(\mathrm{X}',\mathrm{X}^*) & \mathbf{K}(\mathrm{X}',\mathrm{X}') \end{bmatrix}$$

Then components of the GLS equation become:

$$\mathbf{M}\mathbf{m}^{\mathrm{est}} = \bar{\mathbf{f}}^* \quad \text{and} \quad \mathbf{G}\langle\mathbf{m}\rangle = \boldsymbol{\mu}^* \quad \text{and} \quad \mathbf{G}[\mathrm{cov}\,\mathbf{m}]_{\mathrm{A}}\mathbf{G}^{\mathrm{T}} = \mathbf{K}(\mathrm{X}',\mathrm{X}')$$

$$\mathbf{M}[\mathrm{cov}\,\mathbf{m}]_{\mathrm{A}}\mathbf{G}^{\mathrm{T}} = [\mathbf{I} \quad \mathbf{0}] \begin{bmatrix} \mathbf{K}(\mathrm{X}^*,\mathrm{X}^*) & \mathbf{K}(\mathrm{X}^*,\mathrm{X}') \\ \mathbf{K}(\mathrm{X}',\mathrm{X}^*) & \mathbf{K}(\mathrm{X}',\mathrm{X}') \end{bmatrix} \begin{bmatrix} \mathbf{0} \\ \mathbf{I} \end{bmatrix} = [\mathbf{I} \quad \mathbf{0}] \begin{bmatrix} \mathbf{K}(\mathrm{X}^*,\mathrm{X}') \\ \mathbf{K}(\mathrm{X}',\mathrm{X}') \end{bmatrix} = \mathbf{K}(\mathrm{X}^*,\mathrm{X}')$$

Now, the GLS equation exactly matches the GPE equation:

$$\bar{\mathbf{f}}^* = \mathbf{K}(\mathrm{X}^*,\mathrm{X}') + \mathbf{K}(\mathrm{X}^*,\mathrm{X}')\,\{\mathbf{K}(\mathrm{X}^*,\mathrm{X}') + \sigma^2\mathbf{I}\}^{-\mathbf{1}}(\mathbf{y} - \boldsymbol{\mu})$$

(after recognizing that a primed variable in the equation above is the same as an un-starred variable in the GPE equation).

The equivalence is no surprise to me, for both methods use Bayes' theorem to implement prior information and both are based on Normal distributions.

Part 2. Existence of the solution. We consider the case where all model parameters are observed, so that the generalized inverse $\mathbf{G}^{-\mathrm{g}}$ is the $M \times M$ matrix:

$$\mathbf{G}^{-\mathrm{g}} = \gamma^2\mathbf{C}\left[\gamma^2\mathbf{C} + \sigma^2\mathbf{I}\right]^{-1} = \mathbf{C}\left[\mathbf{C} + \beta^2\mathbf{I}\right]^{-1} \quad \text{with} \quad \beta^2 = \frac{\sigma^2}{\gamma^2}$$

Here $\gamma^2\mathbf{C}$ is the signal covariance matrix, with $C_{ii} = 1$, so that γ^2 represents variance. The parameter σ^2 is the variance of uncorrelated noise, so that β^{-2} is the mean-squared signal-to-noise ratio.

We first note some properties of $\mathbf{C}$: (A) Because $\mathbf{C}$ is a normalized covariance matrix, its diagonal elements must be non-negative in all coordinate systems. Consequently, its eigenvalues λ_i are non-negative. (B) Since the quantity $\mathrm{tr}(\mathbf{C})/M = 1$ is invariant under coordinate rotations, the eigenvalues λ_i must "straddle" unity; that is, except for the special case of $\mathbf{C} = \mathbf{I}$, where all eigenvalues $\lambda_i = 1$, some eigenvalues will satisfy $\lambda_i > 1$ and others $\lambda_i < 1$. We will assume that the λ_is are arranged in descending order.

The matrices $\mathbf{C}$ and $[\mathbf{C} + \varepsilon^2\mathbf{I}]$ are simultaneously diagonalizable by the rotation associated with the eigenvector matrix $\mathbf{V}$ of $\mathbf{C}$. Consequently, $\mathbf{G}^{-\mathrm{g}} = \mathbf{V}\mathbf{D}\mathbf{V}^{\mathrm{T}}$, where $\mathbf{D}$ is a diagonal matrix with elements:

$$D_{ii} = \frac{\lambda_i}{\lambda_i + \beta^2}$$

For non-zero values of β^2, the denominator is never zero, since $\lambda_i \geq 0$. Furthermore,

$$D_{ii}(\lambda_i = 0) = 0 \quad \text{and} \quad \lim_{\substack{\varepsilon^2 \to 0 \\ \lambda_i > 0}} D_{ii} = 1 \quad \text{and} \quad \lim_{\varepsilon^2 \to \infty} D_{ii} = 0$$

The limit when $\beta^2 \to 0$, $\lambda_i = 0$ is a removable singularity with a value of unity. Consequently, $0 \le D_{ii} \le 1$ (Figure 1). The matrix $\mathbf{G}^{-g}$ exists and is well-behaved.

Values near these extremes can be calculated with perturbation theory:

$$\text{if} \quad \beta^2 \ll \lambda_i \quad \text{then} \quad D_{ii} = \frac{1}{1+\beta^2/\lambda_i} \approx 1 - \beta^2/\lambda_i$$

$$\text{if} \quad \beta^2 \gg \lambda_i \quad \text{then} \quad D_{ii} = \frac{\lambda_i/\beta^2}{1+\lambda_i/\beta^2} \approx (\lambda_i/\beta^2)(1-\lambda_i/\beta^2) = \lambda_i/\beta^2 - \lambda_i^2/\beta^4$$

Figure 1. Exemplary behavior of D_{ii} in the case where some λ_is are much larger than β^2 and others are much smaller. (A) Plot of eigenvalues $\ln \lambda_i$ versus indices i. (B) Plot of D_{ii}. Note that D_{ii} is between zero and unity.

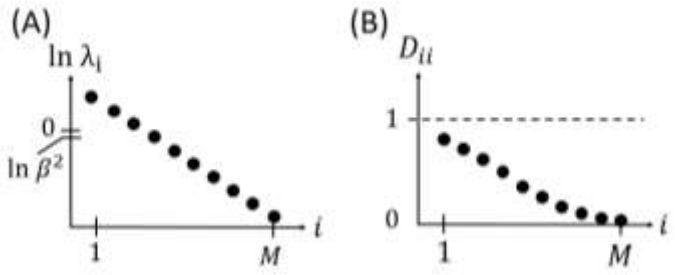

Figure 1. Exemplary behavior of D_{ii} in the case where some λ_is are much larger than β^2 and others are much smaller. (A) Plot of eigenvalues $\ln \lambda_i$ versus indices i. (B) Plot of D_{ii}. Note that D_{ii} is between zero and unity.

Part 3. Impulse Response in the Continuum Limit.

The GLS equation has two parts:

$$\mathbf{m} = \mathbf{C}\boldsymbol{\lambda} \quad \text{with} \quad [\sigma^2\mathbf{I} + \mathbf{C}]\,\boldsymbol{\lambda} = \mathbf{d} \quad \text{and} \quad [\text{cov}\,\mathbf{m}]_A = \gamma^2\mathbf{C}$$

Here, we assume that the matrix $\mathbf{C}$ has a main diagonal of unity, so that the variance at zero lag is γ^2. We now take the continuum limit, with $\mathbf{m}$ becoming $m(x)$, $\boldsymbol{\lambda}$ becoming $\lambda(x)$, $\mathbf{d}$ becoming $d(x)$, and $\gamma^2\mathbf{C}$ becoming $\gamma^2(c(x)*)$:

$$m(x) = \gamma^2 c(x) * \lambda(x) \quad \text{with} \quad \sigma^2\lambda(x) + \gamma^2 c(x) * \lambda(x) = d(x)$$

Here, the normalized auto-covariance function $c(x)$ is unity for zero lag. Note for the low-noise case, $\sigma^2 \ll \gamma^2$ and $m(x) = d(x)$, whereas for the high-noise case, $\gamma^2 \ll \sigma^2$ and $m(x) = (\gamma/\sigma)^2 c(x) * d(x)$. The solution for arbitrary σ^2 can be found by Fourier transforming position x to wavenumber k. Consider the impulse response case, were $d(x) = \delta(x)$:

$$m(k) = \gamma^2 c(k)\lambda(k) \quad \text{and} \quad \lambda(k) = \frac{1}{\gamma^2 c(k) + \sigma^2}$$

$$\text{so} \quad m(k) = \frac{\gamma^2 c(k)}{\gamma^2 c(k) + \sigma^2} = \frac{c(k)}{c(k) + \beta^2} \quad \text{with} \quad \beta^2 = \frac{\sigma^2}{\gamma^2}$$

For the moderate-noise case, with $\sigma^2 > \gamma^2$ and $\beta^2 > 1$, we expand $m(k)$ in a Taylor series:

$$m(k) = \frac{\beta^{-2}c(k)}{[1+\beta^{-2}c(k)]} = \beta^{-2}c(k)[1-\beta^{-2}c(k)+\beta^{-4}c^2(k)-\beta^{-6}c^3(k)+\cdots] =$$

$$= \beta^{-2}c(k)-\beta^{-4}c^2(k)+\beta^{-6}c^3(k)-\beta^{-8}c^4(k)+\cdots$$

Now take inverse Fourier transform:

$$m(x) = \beta^{-2}c(x)-\beta^{-4}c(x)*c(x)+\beta^{-6}c(x)*c(x)*c(x)-\cdots$$

Now consider the special case, where $c(x) = g(x,s)$ with a Gaussian function $g(x,s) \equiv \exp\{-x^2/(2s^2)\}$. As is done with probability density functions, we define its half-width s is in terms of its moments with respect to the origin; that is, $s^2 = M_2/M_0$ where M_0 is the zeroth moment (area) and M_2 is its second moment (variance).

Note that $g(x,s) = \sqrt{2\pi s^2}\,N(x,s)$ where $N(x,s)$ is the Normal distribution. Then, from the rule for convolutions of Normal distributions:

$$N(x,s)*N(x,\gamma,s) = N\big(x,\sqrt{2}s\big)$$

we find:

$$\gamma^2c(x)*\gamma^2c(x) = 2\pi s^2\gamma^4N(x,2s) = \frac{2\pi s^2\gamma^4}{\sqrt{2\pi s^2}}g\big(x,\sqrt{2}s\big) = \sqrt{\pi s^2}g\big(x,\sqrt{2}s\big)$$

$$m(x) = \beta^{-2}g(x,s)-\beta^{-4}\sqrt{\pi s^2}g\big(x,\sqrt{2}s\big)+\cdots$$

Note that the second term is a Gaussian that wider than, and has an opposite sign from, the one in the first term. Its effect is to produce side-lobes in the impulse response.

Part 4. Impulse response with a Gaussian auto-covariance function.

We consider the special case where $c(x)$ is the Gaussian function:

$$c(x) = \exp(-\tfrac{1}{2}s^{-2}x^2) \quad \text{and} \quad c(k) = \sqrt{2\pi}s\exp(-\tfrac{1}{2}s^2k^2)$$

where s is half-width. The Fourier transformed solution is:

$$m(k) = \frac{c(k)}{c(k)+\beta^2} = \frac{\exp(-\tfrac{1}{2}k^2s^2)}{\exp(-\tfrac{1}{2}k^2s^2)+\eta^2} = \frac{1}{1+\eta^2\exp(\tfrac{1}{2}k^2s^2)} \quad \text{with} \quad \eta^2 = \frac{\beta^2}{\sqrt{2\pi}s}$$

Note that since $c(x)$ is a real, symmetric function, so is $c(k)$. The Fourier transformed solution is well behaved, in the sense that:

$$\lim_{k\to 0} m(k) = \frac{1}{1+\eta^2} \equiv A \quad \text{and} \quad \lim_{k\to\pm\infty} m(k) = 0$$

The zero-wavenumber limit yields the area A beneath $m(x)$, which satisfies $A \le 1$. For small values of η^2, the area $A \approx 1$, and for large values of η^2, $A \approx \eta^{-2}$.

Although I have not been able to invert $m(k)$ to $m(x)$, I offer the following analysis of its behavior. First, write:

$$m(k) = a(k)\, b(k) \quad \text{with} \quad a(k) = \frac{1}{\eta^2 + b(k)} \quad \text{and} \quad b(k) = \exp(-½s^2k^2)$$

so that $m(x) = a(x) * b(x)$. Note that $b(k)$ has a maximum of $b_{max} = 1$ at the origin and monotonically decreases with k. Consequently, $a(k)$ monotonically increases with k from a minimum of $(\eta^2 + 1)^{-1}$ at the origin to a maximum of $a_{max} = \eta^{-2}$ at infinity. Now write $a(k) = \eta^{-2}[1 - f(k)]$, with:

$$f(k) = 1 - \eta^2 a(k) = \frac{b(k)}{\eta^2 + b(k)}$$

The "bump function" $f(k)$ (Figure 2) is everywhere non-zero, has a maximum value of $f_{max} = 1/(\eta^2 + 1)$ at the origin, and monotonically decreases towards zero as $k \to \infty$. While we do not calculate their values, the zeroth and second moment, $M_0[f(k)]$ and $M_2[f(k)]$, are positive. Consequently, by the entropic uncertainty principle, $M_2[f(x)] > 0$, too. The zeroth moment $M_o[f(x)] = f_{max}$, and $f(x = 0) = M_0[f(k)] > 0$, since the since the area under a function is the zero-wavenumber value of its Fourier transform and vice versa.

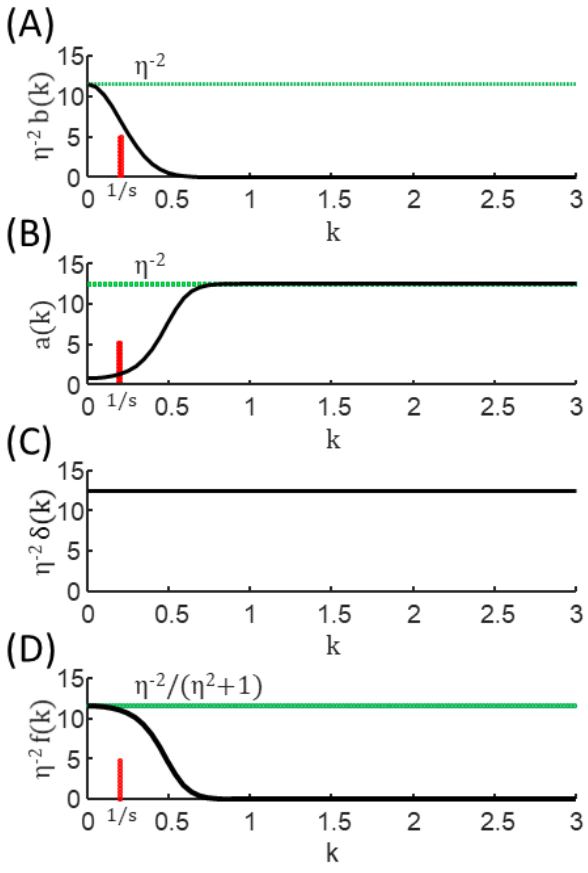

Figure 2. In our analysis of the Gaussian case, the Fourier transform of solution is decomposed into a several functions. (A) The Gaussian function $\eta^{-2}g(k)$ (black curve). (B) The function $a(k)$ (black curve). (C) The function $\eta^{-2}\delta(k)$ (black curve). (D) The "bump" function $\eta^{-2}f(k)$ (black curve). Maximum values (green line) and the point $k = 1/s$ (red bar) are shown. This example uses $\beta^2 = 1$ and $s^2 = 25$. See text for further discussion.

After taking the inverse Fourier transform, the solution is found to be:

$$m(x) = \eta^{-2}\{b(x) - b(x) * f(x)\}$$

Consequently, $\eta^2 m(x)$ consist of the difference between two terms, each of which is positive at $t = 0$. The first term is the Gaussian $b(x)$, which has unit area and half-width s.. The second term is a convolution with area $(\eta^2 + 1)^{-1} \leq 1$, since by moment-convolution theorem $M_0[b(x) * f(x)] = M_0[b(x)]\, M_0[f(x)]$. By the moment-convolution theorem for symmetric functions, the half-width of the second term satisfies:

$$h^2 = \frac{M_2[b(x) * f(x)]}{M_0[b(x) * f(x)]} = \frac{M_2[b(x)]}{M_0[b(x)]} + \frac{M_2[b(x)]}{M_0[b(x)]} > \frac{M_2[b(x)]}{M_0[b(x)]} = s^2$$

Consequently, the second term is wider than the first. The solution $m(x)$ is a function with positive area, a positive central peak and at least one set of side-lobes.

The Fourier transform can also be inverted by numerical means (Figure 3).

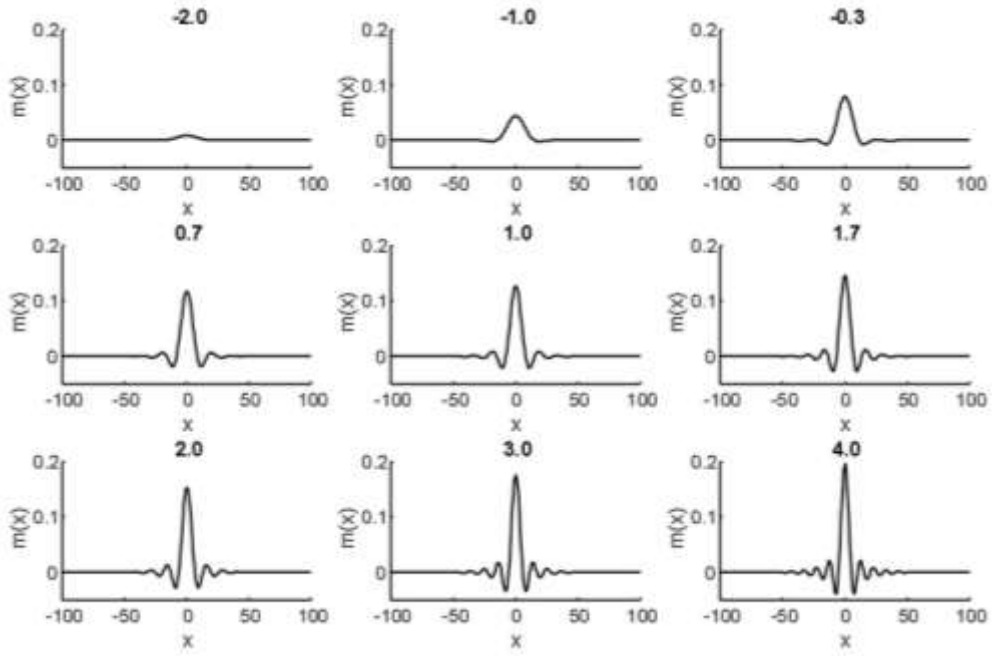

Figure 3. Impulse response for the Gaussian case for variance $s^2 = (8)^2$ and for various values of the logarithmic signal-to-noise ratio $\log_{10}(\beta^{-2})$ (number above graph). Note large sidelobes at the larger values of the ratio.

Part 5. Impulse response with an Exponential auto-covariance function.

We consider the special case of exponential auto-covariance:

$$c(x) = \exp(-\alpha|x|) \quad \text{and} \quad c(k) = \frac{2\alpha}{k^2+\alpha^2} \quad \text{and} \quad \alpha \text{ (units of 1/length)}$$

Using the relation:

$$c(k) + \beta^2 = \frac{2\alpha}{k^2+\alpha^2} + \frac{\beta^2k^2+\beta^2\alpha^2}{k^2+\alpha^2} = \frac{\beta^2k^2+(2\alpha+\beta^2\alpha^2)}{k^2+\alpha^2}$$

We find that:

$$m(k) = \frac{c(k)}{c(k)+\beta^2} = \frac{2\alpha}{k^2+\alpha^2} \times \frac{k^2+\alpha^2}{\beta^2k^2+(2\alpha+\beta^2\alpha^2)} = \frac{2\alpha}{\beta^2k^2+(2\alpha+\beta^2\alpha^2)} =$$

$$= \frac{\left(\frac{\alpha}{\beta^2}\right)\left(\left(\frac{2\alpha}{\beta^2}\right)+\alpha^2\right)^{-½} 2\left(\left(\frac{2\alpha}{\beta^2}\right)+\alpha^2\right)^{½}}{k^2+\left(\left(\frac{2\alpha}{\beta^2}\right)+\alpha^2\right)} =$$

$$= \left(\frac{\alpha}{\rho\beta^2}\right)\frac{2\rho}{k^2+\rho^2} \quad \text{with } \rho = \left((2\alpha/\beta^2)+\alpha^2\right)^{½}$$

Inverting the Fourier Transform yields:

$$m(x) = \frac{\alpha}{\rho\beta^2}\exp(-\rho|x|)$$

The decay rate $\rho \geq \alpha$ and that $\rho \to \alpha$ as $\alpha\beta^2 \to 0$. The area A under $m(x)$ is:

$$A = \int_{-\infty}^{+\infty} m(x)\,dx = \frac{2\alpha}{\rho\beta^2}\int_0^{+\infty}\exp(-\rho x)\,dx = \frac{-2\alpha}{\rho^2\beta^2}\exp(-\rho x)|_0^{+\infty} = \frac{2\alpha}{\rho^2\beta^2} =$$

$$= \frac{2\alpha}{2\alpha+\alpha^2\beta^2} = \frac{1}{1+½\alpha\beta^2}$$

Note that A decreases as $\alpha\beta^2$ increases and that $A \to 1$ as $\alpha\beta^2 \to 0$ (Figure 4).

References

Menke, W., Geophysical Data Analysis: Discrete Inverse Theory, Fourth Edition (Textbook), Elsevier, pp 350, 2018, ISBN: 9780128135556.

Rasmussen, C. E., & Williams, C. K. I., Gaussian processes for machine learning (2016), The MIT Press

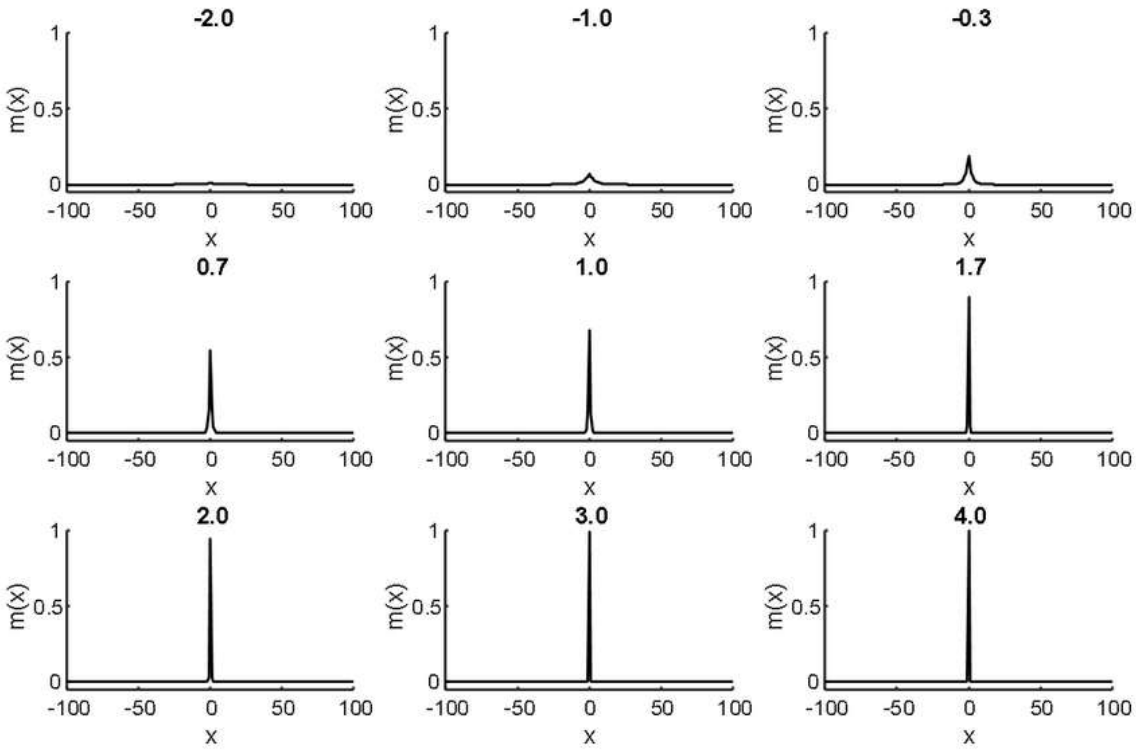

Figure 4. Impulse response for the Exponential case for decay rate $\alpha = \sqrt{2}/s$ corresponding to variance $s^2 = (8)^2$ and for various values of the logarithmic signal-to-noise ratio $\log_{10}(\beta^{-2})$ (number above graph). Note absence of sidelobes.