

ANOVA decomposition of a Gaussian Process model

November 18, 2020

1 ANOVA decomposition of a general function

Functional ANOVA decomposition represents a high-dimensional function as a function of the form

$$f(x_1, x_2, \dots, x_D) = f_0 + \sum_{i=1}^D f_i(x_i) + \sum_{i < j} f_{ij}(x_i, x_j) + \sum_{i < j < k} f_{ijk}(x_i, x_j, x_k) + \dots + f_{1,\dots,D}(x_1, \dots, x_D).$$

Let's first talk about how to compute the ANOVA components in general and then we focus on how to do it for $f(x_1, \dots, x_D)$ which is evaluated as a mean of a Gaussian process.

The first step is to choose the projection operator P :

$$Pf := \int_{[a,b]} f(x) d\mu(x)$$

We are going to use the projection operator P using Lebesgue measure $Pf := \int_{[a,b]} f(x) dx$, so all integrals should work out as expected.

1.1 Two-dimensional case

Now we use the projection operator to define the constant and the main effects. We assume $D = 2$ for now and then introduce some more notation to generalize:

$$f_0 = \int_{[a_1, b_1]} \int_{[a_2, b_2]} f(x_1, x_2) dx_1 dx_2 \tag{1}$$

$$f_1(x_1) = \int_{[a_2, b_2]} \left(f(x_1, x_2) - f_0 \right) dx_2 \tag{2}$$

$$f_2(x_2) = \int_{[a_1, b_1]} \left(f(x_1, x_2) - f_0 \right) dx_1 \tag{3}$$

The interaction effect $f_{1,2}(x_1, x_2)$ is defined as the remainder to make the ANOVA decomposition to work out correctly:

$$f_{1,2}(x_1, x_2) = f(x_1, x_2) - f_0 - f_1(x_1) - f_2(x_2).$$

The **total variance** (TV) of the predictor is defined as

$$\sigma^2(f) := \int_{[a_1, b_1]} \int_{[a_2, b_2]} (f(x_1, x_2) - f_0)^2 dx_1 dx_2$$

One can show that TV is decomposable into the sum of variances of main effects and interactions defined above:

$$\sigma_1^2(f_1) := \int_{[a_1, b_1]} (f_1(x_1))^2 dx_1 \quad (4)$$

$$\sigma_2^2(f_2) := \int_{[a_2, b_2]} (f_2(x_2))^2 dx_2 \quad (5)$$

$$\sigma_{1,2}^2(f_{1,2}) := \int_{[a_1, b_1]} \int_{[a_2, b_2]} (f_{1,2}(x_1, x_2) - f_0 - [f_1(x_1) - f_0] - [f_2(x_2) - f_0])^2 dx_1 dx_2 \quad (6)$$

$$\sigma^2(f) = \sigma_1^2(f_1) + \sigma_2^2(f_2) + \sigma_{1,2}^2(f_{1,2}). \quad (7)$$

And so, by dividing individual variances by TV we can express these components as percentages.

1.2 General case

Using subsets $u \subseteq \{1, \dots, D\}$, we can establish a shorthand notation for ANOVA components, where f_u and x_u represents a subset of vector x with components $x_i, i \in u$. Then we have

$$f(x_1, \dots, x_D) = \sum_{u \subseteq \{1, \dots, D\}} f_u(x_u), \quad (8)$$

$$f_u(x_u) = \int_{\Omega^{D-|u|}} \left(f(x) - \sum_{v \subsetneq u} f_v(x_v) \right) dx_{-u} \quad (9)$$

$$\sigma^2(f_u) = \int_{\Omega^D} \left(f_u(x_u) \right)^2 dx \quad (10)$$

$$\sigma^2(f) = \int_{\Omega^D} \left(f(x) - f_0 \right)^2 dx = \sum_{u \subseteq \{1, \dots, D\}, u \neq \emptyset} \sigma^2(f_u). \quad (11)$$

1.3 ANOVA for Gaussian process

For Bayesian optimization we are using the Gaussian process defined by a parametrized RBF kernel of the form

$$K(x, x' | \theta_0, \vec{\theta}_1) = \theta_0^2 \exp \left(-(x - x')^t D(\vec{\theta}_1) (x - x') \right),$$

where $[D(\vec{\theta}_1)]_{ij} = (\theta_{1,i}^2 + \epsilon) \delta_{ij}$ is a diagonal matrix of scaling coefficients, ϵ being a constant and δ_{ij} being Dirac-delta.

$$K(x, x' | \theta_0, \vec{\theta}_1) = \theta_0^2 \exp \left(- \sum_{d=1}^D (\theta_{1,d}^2 + \epsilon) (x_d - x'_d)^2 \right) = \theta_0^2 \prod_{d=1}^D \exp \left(- (\theta_{1,d}^2 + \epsilon) (x_d - x'_d)^2 \right).$$

I drop θ 's from the definition of K for shorthand.

Let Σ be the correlation matrix of the training set $\{(x_i, y_i)\}_{i=1}^n$ and $k(x)$ a vector with correlations to the new point x , to evaluate the functional given by the GP and get the mean at the points x we do:

$$\Sigma := \{K(x_i, x_j)\}_{ij} \quad (12)$$

$$k(x) := \{K(x, x_i)\}_{i=1}^n \quad (13)$$

$$f(x) := k(x)^t \Sigma^{-1} y \quad (14)$$

$$= \sum_{i=1}^n K(x, x_i) \Sigma_i^{-1} y, \quad (15)$$

where Σ_i^{-1} is the i 'th row of the inverse correlation matrix (of course, we are just solving $\Sigma z = y$ and take the i 'th component of z , so I will just use z_i as a shorthand:

$$f(x) = \sum_{i=1}^n z_i K(x, x_i)$$

I'll use wolframalpha to derive integrands. We also define a shorthand

$$c_d := \sqrt{\theta_{1d}^2 + \epsilon}$$

For the [constant](#) we have:

$$f_0 = \int_{\Omega_D} f(x) dx = \int_{\Omega_D} \sum_{i=1}^n z_i K(x, x_i) dx \quad (16)$$

$$= \sum_{i=1}^n z_i \theta_0^2 \prod_{d=1}^D \int_{a_d}^{b_d} \exp(-c_d^2 (x_d - x_{id})^2) dx_d \quad (17)$$

$$= \theta_0^2 \sum_{i=1}^n z_i \prod_{d=1}^D \frac{\sqrt{\pi} [\operatorname{erf}(c_d (b_d - x_{id})) - \operatorname{erf}(c_d (a_d - x_{id}))]}{2c_d} \quad (18)$$

For the [total variance](#) we have:

$$\sigma^2(f) = \int_{\Omega_d} (f(x) - f_0)^2 dx = \int_{\Omega_d} \left(\sum_{i=1}^n z_i K(x, x_i) - f_0 \right)^2 dx \quad (19)$$

$$= \int_{\Omega_d} \sum_{i=1}^n \sum_{j=1}^n z_i z_j K(x, x_i) K(x, x_j) dx - 2f_0 \int_{\Omega_d} \sum_{i=1}^n z_i K(x, x_i) dx + f_0^2 \int_{\Omega_d} dx \quad (20)$$

$$= f_0^2 \left(\prod_{d=1}^D (b_d - a_d) - 2 \right) + \theta_0^4 \sum_{i=1}^n \sum_{j=1}^n z_i z_j \prod_{d=1}^D \int_{a_d}^{b_d} \exp \left(-c_d^2 \left[(x_d - x_{id})^2 + (x_d - x_{jd})^2 \right] \right) dx_d \quad (21)$$

$$= f_0^2(\dots) + \theta_0^4 \sum_{i=1}^n \sum_{j=1}^n z_i z_j \prod_{d=1}^D \frac{\sqrt{\frac{\pi}{2}} \exp \left(-1/2 c_d^2 (x_{id} - x_{jd})^2 \right) \left(\operatorname{erf} \left[\frac{c_d(-2a_d + x_{id} + x_{jd})}{\sqrt{2}} \right] - \operatorname{erf} \left[\frac{c_d(-2b_d + x_{id} + x_{jd})}{\sqrt{2}} \right] \right)}{2c_d} \quad (22)$$

For the main effects we have:

$$f_t(x_t) = \theta_0^2 \sum_{i=1}^n z_i \exp \left(-c_t^2 (x_t - x_{it})^2 \right) \prod_{d \neq t}^D \frac{\sqrt{\pi} \left[\operatorname{erf} \left(c_d (b_d - x_{id}) \right) - \operatorname{erf} \left(c_d (a_d - x_{id}) \right) \right]}{2c_d} - f_0 \prod_{d=1}^D (b_d - a_d) \quad (23)$$

$$=: \theta_0^2 \sum_{i=1}^n z_i \exp \left(-c_t^2 (x_t - x_{it})^2 \right) P_i - F_0 \quad (24)$$

$$\sigma^2(f_t(x_t)) = c \left(f_d(x_d) \right)^2 dx = \int_{\Omega^D} \left(\theta_0^2 \sum_{i=1}^n z_i \exp \left(-c_t^2 (x_t - x_{it})^2 \right) P_i - F_0 \right)^2 dx \quad (25)$$

$$= \int_{\Omega^D} \theta_0^4 \sum_{i=1}^n \sum_{j=1}^n z_i z_j P_i P_j \exp \left(-c_t \left[(x_t - x_{it})^2 + (x_t - x_{jt})^2 \right] \right) \quad (26)$$

$$- 2\theta_0^2 F_0 \sum_{i=1}^n z_i P_i \int_{\Omega^D} \exp \left(-c_t^2 (x_t - x_{it})^2 \right) dx + F_0^2 \int_{\Omega^D} dx \quad (27)$$

$$= \theta_0^4 \sum_{i=1}^n \sum_{j=1}^n z_i z_j P_i P_j \frac{\sqrt{\frac{\pi}{2}} e^{-1/2 c_t^2 (x_{it} - x_{jt})^2} \left(\operatorname{erf} \left[\frac{c_t(-2a_t + x_{it} + x_{jt})}{\sqrt{2}} \right] - \operatorname{erf} \left[\frac{c_t(-2b_t + x_{it} + x_{jt})}{\sqrt{2}} \right] \right)}{2c_t} \prod_{d \neq t}^D (b_d - a_d) \quad (28)$$

$$- 2\theta_0^2 F_0 \sum_{i=1}^n z_i P_i \frac{\sqrt{\pi} \left[\operatorname{erf} \left(c_t (b_t - x_{it}) \right) - \operatorname{erf} \left(c_t (a_t - x_{it}) \right) \right]}{2c_t} \prod_{d \neq t}^D (b_d - a_d) + F_0^2 \prod_{d=1}^D (b_d - a_d) \quad (29)$$

Note: we can further simplify the computations if we normalize the domain Ω^D to a hypercube $[0, 1]^D$.