Kriging

In my view Kriging can be summarize as random field prediction in the presence of functional covariates with unknown coefficients.

"classic" model where Kriging is applied.

Data:

$d_1, \dots, d_n$ at spatial locations $x_1, \dots, x_n \in \mathbb{R}^d$.

Model:

(K1) $d_i = Y(x_i) + \varepsilon_i$

(K2) $Y(x) = \sum_{p=1}^{m} \beta_p f_p(x) + Z(x)$

(K3) $f_p : \mathbb{R}^d \to \mathbb{R}$ are know "functional covariates".

(K4) β_p are unknown coefficients

(K5) $(\varepsilon_1, \dots, \varepsilon_n) \sim N(0, \sigma^2 I)$.

(K6) $Z \sim GRF_{\mathbb{R}^d}(0, K)$, $K : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$.

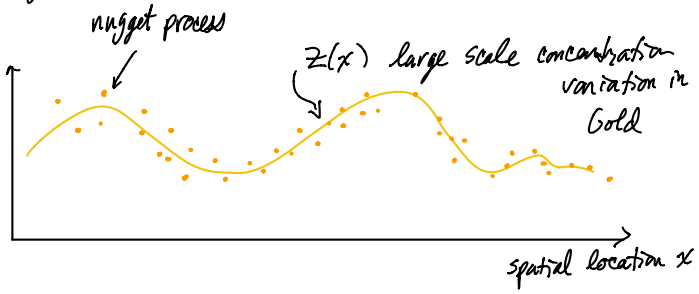
Goal:

Predict $Y(x_0)$ at some $x_0 \in \mathbb{R}^d$

This is a natural model in the sense that the linear model can explain large scale (deterministic) variation, ε_i's account for spatially uncorrelated error and $Z(x)$ accounts for "unknown spatial covariates" which induce spatially correlated errors using the linear model $\sum_p \beta_p f_p(x)$.

Note: The additive errors $\varepsilon_1, \dots, \varepsilon_n$ correspond to what is called a nugget effect. It can be used to model instrumental noise or a microscale process that has correlation length scale much smaller than $Z(x)$.

The name comes from mining applications e.g.



There are a few equivalent characterizations of Kriging, each giving a different perspective but yeild equivalent predictions.

Kriging with Generalized least square Regression

I think this is the most intuitive and direct but makes it difficult to see why one can use generalized cov funs (later).

Note that

$$\begin{bmatrix} d_1 \\ \vdots \\ d_n \\ Y(x_0) \end{bmatrix} \equiv \begin{bmatrix} \sum_p \beta_p f_p(x_1) \\ \vdots \\ \sum_p \beta_p f_p(x_n) \\ \sum_p \beta_p f_p(x_0) \end{bmatrix} + \begin{bmatrix} Z(x_1) \\ \vdots \\ Z(x_n) \\ Z(x_0) \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \\ 0 \end{bmatrix}$$

So letting $f(x_0) = \begin{bmatrix} f_1(x_0) \\ \vdots \\ f_m(x_0) \end{bmatrix}$ & $f(x_1,\dots,x_n) = \begin{bmatrix} | & & | \\ f(x_1) & \cdots & f(x_n) \\ | & & | \end{bmatrix}$

we have

$$\begin{bmatrix} d_1 \\ \vdots \\ d_n \\ \hline Y(x_0) \end{bmatrix} \sim N\left(\underbrace{\begin{bmatrix} \vdots \\ f(x_1,\dots,x_n)^T\beta \\ \vdots \\ \hline f(x_0)^T\beta \end{bmatrix}}_{\| \; \begin{bmatrix} X\beta \\ \hline f(x_0)^T\beta \end{bmatrix}}, \underbrace{\left[\begin{array}{ccc|c} & [K(x_i,x_j)]_{i,j=1}^n + \sigma^2 I & & \begin{matrix} K(x_1,x_0) \\ \vdots \\ K(x_n,x_0) \end{matrix} \\ \hline \cdots & K(x_0,x_j) & \cdots & K(x_0,x_0) \end{array}\right]}_{\| \; \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{10} \\ \hline \Sigma_{01} & \Sigma_{00} \end{array}\right]} \right)$$

So $d = X\beta + \varepsilon$ is simply a GLS regression problem:

$$\hat{\beta} = (X^T \Sigma_{11}^{-1} X)^{-1} X^T \Sigma_{11}^{-1} d \qquad (1a)$$

To predict $Y(x_0)$ we simply use the multivariate model for $(d^T, Y(x_0))^T$ as above and write

$$E(Y(x_0)|d) = \Sigma_{01}\Sigma_{11}^{-1}(d - X\beta) + f(x_0)^T\beta$$

which requires known β so the Kriging prediction simply plugs in $\hat{\beta}$:

$$\hat{Y}(x_0) := \Sigma_{01}\Sigma_{11}^{-1}(d - X\hat{\beta}) + f(x_0)^T\hat{\beta} \qquad (1b)$$

At this point its easy to remember and compute but its not clear the impact of plugging in $\hat{\beta}$ for β in $E(Y(x_0)|d)$.

The next characterization shows it is optimal among "linear unbiased predictions".

Kriging as a best linear unbiased predictor

The idea predict $\widehat{Y(x_0)}$ via $d^T \cdot \lambda$

where $\lambda = \underset{\xi \in \mathcal{UB}}{\arg\min}\; E\left(Y(x_0) - d^T \cdot \xi\right)^2$

where the "unbias" constraint $\lambda \in \mathcal{UB}$ is:

$$E(\hat{Y}(x_0)) = f(x_0)^T\beta$$

Note the LHS equals

$$E(d^T \cdot \lambda) = \left(f(x_1,\dots,x_n)^T \cdot \beta\right)^T \cdot \lambda$$
$$= \beta^T \cdot f(x_1,\dots,x_n) \cdot \lambda$$

So the unbiased constraint becomes

$$E(\hat{Y}(x_0)) = f(x_0)^T\beta$$
$$\Updownarrow$$
$$\forall \beta \in \mathbb{R}^m, \quad \beta^T \cdot f(x_1,\dots,x_n) \cdot \lambda = \beta^T \cdot f(x_0)$$
$$\Updownarrow$$

$$\underbrace{f(x_1,\dots,x_n)}_{\text{length } n \dots \text{ i.e. the number of obs points}} \cdot \underbrace{\lambda}_{n\times 1} = \underbrace{f(x_0)}_{m\times 1}$$

$$\Updownarrow$$

$$\lambda \in \mathcal{UB}$$

Remark: Notice that if we define

$$\lambda^{ext} = \begin{pmatrix} \lambda_1, \lambda_2, \dots, \lambda_n, -1 \end{pmatrix}^T$$

then $f(x_1, \dots, x_n, x_0) \cdot \lambda^{ext} = 0$

To summarize we predict $\hat{Y}(x_0) = d^T \cdot \lambda$

where λ satisfies

$$f(x_1, \dots, x_n) \cdot \lambda = f(x_0)$$

i.e. $f(x_1, \dots, x_n, x_0) \cdot \lambda^{ext} = 0$

and minimizes

$$E\left(Y(x_0) - d^T \cdot \lambda\right)^2$$

$$= E\left(\begin{bmatrix} d \\ Y(x_0) \end{bmatrix}^T \cdot \lambda^{ext}\right)^2$$

$$= \left(\lambda^{ext}\right)^T \left[\begin{array}{c|c} \left[K(x_i, x_j)\right]_{i,j=1}^n + \sigma^2 I & \begin{matrix} \vdots \\ K(x_j, x_0) \\ \vdots \end{matrix} \\ \hline \cdots K(x_0, x_j) \cdots & K(x_0, x_0) \end{array}\right] \lambda^{ext} - 0$$

$$= \left(\lambda^{ext}\right)^T \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{10} \\ \hline \Sigma_{01} & \Sigma_{00} \end{array}\right] \lambda^{ext}$$

This term is $\left(f(x_1, \dots, x_n, x_0) \cdot \lambda^{ext}\right)^2$ but is zero due to the unbias constraint

---

Recall the technique of using Lagrange multipliers for constraind minimization:

$$\lambda^0 \in \underset{\{\lambda : G(\lambda) = 0\}}{\arg\min}\ J(\lambda)$$

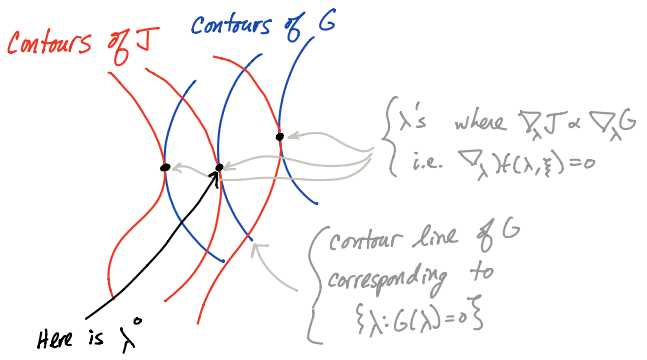
where $\dim(G(\lambda)) = m$, $\dim(\lambda) = n > m$

$$\Updownarrow$$

$$\lambda^0 \in \left\{ \lambda : \exists \xi \in \mathbb{R}^m \text{ s.t. } \begin{matrix} \nabla \mathcal{H}(\lambda, \xi) = 0 \\ \text{and } G(\lambda) = 0 \end{matrix} \right\}$$

with $\mathcal{H}(\lambda, \xi) = J(\lambda) + \xi^T \cdot G(\lambda)$

The picture looks like this



In our case we have

$$\mathcal{H}(\lambda, \xi) = J(\lambda) + \xi^T \cdot G(\lambda)$$

$$J(\lambda): \quad \frac{1}{2}\begin{bmatrix} \lambda \\ -1 \end{bmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{10} \\ \Sigma_{01} & \Sigma_{00} \end{bmatrix} \begin{bmatrix} \lambda \\ -1 \end{bmatrix}$$

$$G(\lambda): \quad \xi^T \cdot \left(f(x_1, \dots, x_n)\lambda - f(x_0)\right)$$

Also

$$\frac{\partial J(\lambda)}{\partial \lambda} = \begin{bmatrix} \lambda \\ -1 \end{bmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{10} \\ \Sigma_{01} & \Sigma_{00} \end{bmatrix} \begin{bmatrix} I_{n \times n} \\ 0_{1 \times n} \end{bmatrix}$$

so that

$$\nabla_\lambda J(\lambda) = \begin{bmatrix} I_{n \times n} & 0_{n \times 1} \end{bmatrix} \begin{bmatrix} \Sigma_{11} & \Sigma_{10} \\ \Sigma_{01} & \Sigma_{00} \end{bmatrix} \begin{bmatrix} \lambda \\ -1 \end{bmatrix}$$

$$= \Sigma_{11}\lambda - \Sigma_{10}$$

and therefore

$$\nabla_\lambda \mathcal{H}(\lambda, \xi) = 0 \text{ and } G(\lambda) = 0$$

$$\Updownarrow$$

$$\Sigma_{11}\lambda - \Sigma_{10} + f(x_1, \dots, x_n)^T \cdot \xi = 0 \text{ and } f(x_1, \dots, x_n)\lambda - f(x_0) = 0$$

$$\Updownarrow$$

$$\begin{bmatrix} \Sigma_{11} & f(x_1, \dots, x_n)^T \\ f(x_1, \dots, x_n) & 0_{m \times m} \end{bmatrix} \begin{bmatrix} \lambda \\ \xi \end{bmatrix} = \begin{bmatrix} \Sigma_{10} \\ f(x_0) \end{bmatrix}$$

Therefore the Kriging predictor via (unbias) constrained minimizing MSE is:

$$\hat{Y}(x_0) = d^T \lambda \tag{2a}$$

where λ satisfies

$$\begin{bmatrix} \Sigma_{11} & f(x_1,\dots,x_n)^T \\ f(x_1,\dots,x_n) & 0_{m\times m} \end{bmatrix} \begin{bmatrix} \lambda \\ \xi \end{bmatrix} = \begin{bmatrix} \Sigma_{10} \\ f(x_0) \end{bmatrix} \tag{2b}$$

with $\Sigma_{11} = [K(x_i,x_j)]_{i,j=1}^n + \sigma^2 I$ and $\Sigma_{10} = \begin{bmatrix} \vdots \\ K(x_j,x_0) \\ \vdots \end{bmatrix}$

Remark: Notice that in this characterization we automatically get an estimate of the variance of prediction error by

$$E\left(Y(x_0) - \hat{Y}(x_0)\right)^2$$

$$= \begin{bmatrix} \lambda \\ -1 \end{bmatrix}^T \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{10} \\ \hline \Sigma_{01} & \Sigma_{00} \end{array}\right] \begin{bmatrix} \lambda \\ -1 \end{bmatrix} \tag{2c}$$

where $(\lambda^T, \xi^T)^T$ solves the above Lagrange multiplier problem.

Kriging as a spline interpolator

A typical spline interpolator has the form:

$$\hat{Y}(x_0) = \sum_{k=1}^{n} c_k K(x_0, x_k) + \sum_{p=1}^{m} b_p f_p(x_0)$$

(K: reproducing kernel; x_k: knots)

where $c = (c_1,\dots,c_n)^T$ & $b = (b_1,\dots,b_m)^T$ are given by minimizing "bending energy" and "data fidelity" as follows

$$(c^*, b^*) \in \underset{(c,b)}{\operatorname{argmin}} \left\{ \mathcal{D}(c,b) + \mathcal{E}(c) \right\}$$

where

$\mathcal{D}(c,b) \equiv$ data fidelity

$$= \frac{1}{2\sigma^2} \left\| d - [\hat{Y}(x_1),\dots,\hat{Y}(x_n)]^T \right\|^2$$

$$= \frac{1}{2\sigma^2} \left\| d - (Kc + F^T b) \right\|^2$$

($K = [K(x_i,x_j)]_{i,j=1}^n$; $F := f(x_1,\dots,x_n)$)

and

$\mathcal{E}(c) \equiv$ bending energy (with $\sum_{p=1}^m b_p f_p$ unpenalized)

$$= \frac{1}{2} c^T K c$$

$$\tfrac{1}{2} \langle Kc | K^{-1} | Kc \rangle$$

Now the minimizer is characterized by

$$\nabla_c \mathcal{D}(c,b) + \nabla_c \mathcal{E}(c) = 0$$

$$\nabla_b \mathcal{D}(c,b) = 0$$

$$\Updownarrow$$

$$-(d - Kc - F^T b) + \sigma^2 c = 0 \quad \leftarrow \text{after a left multiply with } \sigma^2 K^{-1}$$

$$-F(d - Kc - F^T b) = 0$$

$$\Updownarrow \quad \leftarrow \text{this is } \sigma^2 c \text{ from the line above}$$

$$(K + \sigma^2 I)c + F^T b = d$$

$$Fc = 0$$

In summary the Kriging predictor (defined via a spline approach) is defined as:

$$\hat{Y}(x_0) = \sum_{k=1}^{n} c_k K(x_0, x_k) + \sum_{p=1}^{m} b_p f_p(x_0) \quad (3a)$$

where

$$\begin{bmatrix} \Sigma_{11} & f(x_1,\dots,x_n)^T \\ f(x_1,\dots,x_n) & 0_{m\times m} \end{bmatrix} \begin{bmatrix} c \\ b \end{bmatrix} = \begin{bmatrix} d \\ 0 \end{bmatrix} \quad (3b)$$

with $\Sigma_{11} = [K(x_i,x_j)]_{i,j=1}^{n} + \sigma^2 I$.

All three Kriging characterizations are the same

$\hat{Y}(x_0)$ from (2a)

$$= d^T \cdot \lambda$$

$$= \begin{bmatrix} \lambda \\ \xi \end{bmatrix}^T \cdot \begin{bmatrix} d \\ 0 \end{bmatrix}$$

(2b)

$$= \begin{bmatrix} \Sigma_{10} \\ f(x_0) \end{bmatrix}^T \cdot \begin{bmatrix} \Sigma_{11} & f(x_1,\dots,x_n)^T \\ f(x_1,\dots,x_n) & 0_{m\times m} \end{bmatrix}^{-1} \begin{bmatrix} d \\ 0 \end{bmatrix}$$

(3b)

$$= \begin{bmatrix} \Sigma_{10} \\ f(x_0) \end{bmatrix}^T \cdot \begin{bmatrix} c \\ d \end{bmatrix}$$

$$= \hat{Y}(x_0) \text{ from (3a)}$$

$$= \hat{Y}(x_0) \text{ from (1a)}$$

verified by noticing that $b := \hat{\beta}$ and $c = \Sigma^{-1}(d - f(x_1,\dots,x_n)^T\hat{\beta})$ satisfy (3b)

Now

- $\hat{Y}(x_0)$ from (1a) is probably easiest to remember and intuitive
- $\hat{Y}(x_0)$ from (2a) shows this predictor is optimal (in m.s. sense) among all unbiased predictors and gives prediction uncertainty.
- $\hat{Y}(x_0)$ from (3a) is used to establish all three are equal and gives a "bayesian" interpretation (viewing $\mathcal{E}(c)$ as a "prior").

Replacing the nugget variance with a Noise-to-Signal parameter

By examing the Kriging prediction $\hat{Y}(x_0)$ given in the three different forms above, one discovers a number of simplifications that reduce the number of parameters needed. Here is one example.

Start by writing the Kriging model with an overall variance parameter σ_Z^2 for Z and the nugget variance σ_ε^2 for ε_k

$$(*)\quad \begin{cases} d_k = \sum_{p=1}^{m} b_p f_p(x_k) + Z(x_k) + \varepsilon_k \\ Z \sim GRF_{pd}(0, \sigma_Z^2 K) \quad \text{so } K(x,x)=1 \\ \varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} N(0, \sigma_\varepsilon^2) \end{cases}$$

by simply factoring out σ_Z^2 one notices $\hat{Y}(x_0)$ in (1a) from model (*) is the same prediction as $\hat{Y}(x_0)$ in (1a) from model (**):

$$(**)\quad \begin{cases} d_k = \sum_{p=1}^{m} b_p f_p(x_k) + Z(x_k) + \varepsilon_k \\ Z \sim GRF_{pd}(0, K) \quad K(x,x)=1 \text{ still} \\ \varepsilon_1, \dots, \varepsilon_n \overset{iid}{\sim} N(0, \frac{\sigma_\varepsilon^2}{\sigma_Z^2}) \quad \text{ratio of nugget var to } Z \text{ var.} \end{cases}$$

So once the correlation function K and the covariates f_p are defined the behavior of $\hat{Y}(x_0)$ is completely determined by a single parameter: $\sigma_\varepsilon^2/\sigma_Z^2 = \frac{N}{S}$

Warning: To generate error bars for $\hat{Y}(x_0)$ with (2a) and (2b) you will need to know both σ_ε^2 and σ_Z^2 seperately.

Kriging with the variogram

Recall the characterization of $\hat{Y}(x_0)$ given in (2a) and (2b) defined $\hat{Y}(x_0) = d^T \cdot \lambda$ where λ is defined by

$$\underset{\lambda \in \mathbb{R}^B}{\operatorname{argmin}} \; E\left(Y(x_0) - d^T \cdot \lambda\right)^2$$

$$= \underset{\lambda \in \mathbb{R}^B}{\operatorname{argmin}} \begin{bmatrix} \lambda \\ -1 \end{bmatrix}^T \left[\begin{array}{c|c} \Sigma_{11} & \Sigma_{10} \\ \hline \Sigma_{01} & \Sigma_{00} \end{array}\right] \begin{bmatrix} \lambda \\ -1 \end{bmatrix}$$

$$= \underset{\lambda \in \mathbb{R}^B}{\operatorname{argmin}} \left\{ \begin{bmatrix} \lambda \\ -1 \end{bmatrix}^T \left(K(z_i, z_j)\right)_{i,j=1}^{n+1} \begin{bmatrix} \lambda \\ -1 \end{bmatrix} + \sigma_\varepsilon^2 \|\lambda\|_2^2 \right\}$$

where $(z_1, \dots, z_{n+1}) = (x_1, \dots, x_n, x_0)$

Also recall that

$$\lambda \in \mathbb{R}^B \iff f(x_1, \dots, x_n, x_0) \cdot \begin{bmatrix} \lambda \\ -1 \end{bmatrix} = 0$$

$$\iff f(z_1, \dots, z_{n+1}) \begin{bmatrix} \lambda \\ -1 \end{bmatrix} = 0$$

Therefore we get the same λ, i.e. same $\hat{Y}(x_0)$ and predicted sd, if we replace $K(x,y)$ with $\tilde{K}(x,y)$ so long as the following holds:

$$\begin{bmatrix} \lambda \\ -1 \end{bmatrix}^T \left(K(z_i, z_j)\right)_{i,j=1}^{n+1} \begin{bmatrix} \lambda \\ -1 \end{bmatrix} = \begin{bmatrix} \lambda \\ -1 \end{bmatrix}^T \left(\tilde{K}(z_i, z_j)\right)_{i,j=1}^{n+1} \begin{bmatrix} \lambda \\ -1 \end{bmatrix}$$

whenever $f(z_1, \dots, z_{n+1}) \begin{bmatrix} \lambda \\ -1 \end{bmatrix} = 0$

Or more generally if $\forall n$, $\forall y_1, \dots, y_n \in \mathbb{R}^q$, and $\forall c_1, \dots, c_n \in \mathbb{R}$ we have

$$c^T \left(K(y_i, y_j)\right)_{i,j=1}^n c = c^T \left(\tilde{K}(y_i, y_j)\right)_{i,j=1}^n c$$

whenever $f(y_1, \dots, y_n)^T \cdot c = 0$

Example

Suppose $m = 1$ and the only covariate we have is constant, i.e

$$f_1(x) \equiv 1.$$

Then

$$f(y_1, \dots, y_n) \cdot c = 0 \iff c_1 + \cdots + c_n = 0$$

Now consider cov fun $K(x,y)$ and the variogram of $K(x,y)$

$$\gamma(x,y) = K(x,x) + K(y,y) - 2K(x,y)$$

$$K(x,y) = \frac{1}{2}\left(K(x,x) + K(y,y) - \gamma(x,y)\right)$$

$$\tilde{K}(x,y) = \frac{1}{2}\left(\gamma(x,a_0) + \gamma(y,a_0) - \gamma(x,y)\right)$$

If $c_1 + \cdots + c_n = 0$ then

$$\sum_{k,\rho=1}^n c_k c_\rho K(y_k, y_\rho)$$

$$= \frac{1}{2} \sum_{k,\rho=1}^n c_k c_\rho K(y_k, y_k) \quad \text{(Both } y_k\text{)}$$

$$+ \frac{1}{2} \sum_{k,\rho=1}^n c_k c_\rho K(y_\rho, y_\rho) \quad \text{(Both } y_\rho\text{)}$$

$$- \frac{1}{2} \sum_{k,\rho=1}^n c_k c_\rho \gamma(y_k, y_\rho)$$

But since

$$\frac{1}{2}\sum_{k,\rho=1}^{n} c_k c_\rho K(y_k, y_k)$$

$$= \frac{1}{2}\Big(\underbrace{\sum_{\rho=1}^{n} c_\rho}_{=0}\Big)\Big(\sum_{k=1}^{n} c_k K(y_k, y_\rho)\Big)$$

$$= 0$$

and similarly $\frac{1}{2}\sum_{k,\rho=1}^{n} c_k c_\rho K(y_\rho, y_\rho) = 0$

we have

$$\sum_{k,\rho=1}^{n} c_k c_\rho \, K(y_k, y_\rho) = \sum_{k,\rho=1}^{n} c_k c_\rho \, \frac{(-1)}{2}\gamma(y_k, y_\rho)$$

$$= \sum_{k,\rho=1}^{n} c_k c_\rho \, \tilde{K}(y_k, y_\rho)$$

↑ same reasoning

whenever $c_1 + \cdots + c_n = 0$.

i.e. We just need to use the $\frac{(-1)}{2}$ variogram in place of $K(x,y)$ for Kriging and s.d. prediction, when one of the covariates $f_\rho(x) \equiv 1$.

Notice also that if one has scale and variance parameters (ρ, σ^2)... and use the variogram $\gamma(x,y) = \|x-y\|^{2\nu}$ (with $\sigma=1, \rho=1$) then

$$\sigma^2 \gamma\left(\frac{x}{\rho}, \frac{y}{\rho}\right) = \frac{\sigma^2}{\rho^{2\nu}}\gamma(x,y)$$

↑ The separate ρ, σ merge into a single parameter $\sigma^2/\rho^{2\nu}$.

Kriging with a generalized covariance function

It is pretty hard to extend the technique for variograms with $f_1(x) \equiv 1$ to general covariates $f_\rho(x)$ ... with one exception:

$$\left\{\begin{matrix}\text{polynomials} \\ \text{of order} \le q\end{matrix}\right\} \subseteq \text{span}\{f_1(x), \ldots, f_m(x)\}. \quad \text{(i)}$$

For convenience let the monomials of order $\le q$ be denoted

$$\underbrace{m_0(x)}_{=1}, \ldots, m_{M_q}(x)$$

So we seek functions $\tilde{K}(x,y)$ s.t.

$$\left.\begin{matrix} f(y_1, \ldots, y_n)\cdot c = 0 \text{ and (i)} \\ \Downarrow \\ \sum_{k,\rho=1}^{n} c_k c_\rho K(y_k, y_\rho) = \sum_{k,\rho=1}^{n} c_k c_\rho \tilde{K}(y_k, y_\rho) \end{matrix}\right\} \text{(ii)}$$

holds for any $y_1, \ldots, y_n \in \mathbb{R}^d$ and $c_1, \ldots, c_n \in \mathbb{R}$.

A $\tilde{K}(x,y)$ with this property (for some positive definite K) is called a generalized covariance function of order q.

The official (but equivalent) definition is usually given as follows:

Definition: $\mathcal{Y}(x,y): \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a **generalized covariance function on $\mathbb{R}^d$ of order q**

$$\left.\begin{matrix} x_1, \ldots, x_n \in \mathbb{R}^d, c_1, \ldots, c_n \in \mathbb{R} \\ \text{and } \sum_{i=1}^{n} c_i m(x_i) = 0 \\ \text{for all monomials } m(\cdot) \\ \text{on } \mathbb{R}^d \text{ of order} \le q \end{matrix}\right\} \Rightarrow \sum_{i,j=1}^{n} c_i c_j \mathcal{Y}(x_i, x_j) \ge 0$$

Remark: Smaller $q \Rightarrow$ more stringent, in above definition. If $\mathcal{Y}$ satisfies above for q, then also for any $q' > q$.

To find a $\tilde{K} \neq K$ that satisfies (ii) recall how it worked in the previous section where $f_1(x) \equiv 1$. In this case we defined

$$\tilde{Z}(x) := Z(x) - Z(a_0)$$

which removed an overall constant and discovered $\tilde{K}(x,y) = \text{cov}(\tilde{Z}(x), \tilde{Z}(y))$ can be used inplace of $K(x,y)$ to yeild equivalent Kriging predictions.

In the present case this suggests to define $\tilde{Z}$ by removing all monomials of order $\leq q$ from Z so that at some fixed points

$$a_0, a_1, \ldots, a_{M_q} \in \mathbb{R}^d$$

we have $\tilde{Z}(a_i) = 0$.

To this end define

$$\tilde{Z}(x) := Z(x) - \varphi^i(x) Z(a_i)$$

(Short for $\sum_{i=0}^{M_q} \varphi^i(x) Z(a_i)$)

where $\varphi^i(x) Z(a_i)$ is given by

$$\underbrace{\begin{bmatrix} 1 \\ m_1(x) \\ \vdots \\ m_{M_q}(x) \end{bmatrix}^T \begin{bmatrix} 1 & m_1(a_0) & & m_{M_q}(a_0) \\ \vdots & \vdots & & \vdots \\ 1 & m_1(a_N) & & m_{M_q}(a_N) \end{bmatrix}^{-1}}_{= \left[\varphi^0(x), \varphi^1(x), \ldots, \varphi^{M_q}(x)\right]} \begin{bmatrix} Z(a_0) \\ \vdots \\ Z(a_{M_q}) \end{bmatrix}$$

This is what we want since

$$\varphi^i(a_j) Z(a_i) = Z(a_j), \quad \forall j$$

$$\therefore \quad \tilde{Z}(a_j) = 0 \quad \forall j$$

Note: there is an implicit assumption here that the a_i's are "spread out" enough so the above matrix inverse is defined.

Claim: $\tilde{K}(x,y) = \text{cov}(\tilde{Z}(x), \tilde{Z}(y))$ defined as above satisfies (ii).

Proof:

$$\tilde{K}(x,y) = K(x,y) - \varphi^i(x) K(y, a_i) - \varphi^i(y) K(x, a_i) + \varphi^i(x) \varphi^j(y) K(a_i, a_j)$$

which can be written in operator notation

$$\tilde{K}(x,y) = \iint ds\, dt \left[\delta_{x-s} + \varphi^i(x) \delta_{a_i - s}\right]\left[\delta_{y-t} + \varphi^j(y) \delta_{a_j - t}\right] K(s,t)$$

where $\int ds \int_{x-s} f(s) \equiv f(x)$

Now suppose (i) and $f(y_1, \ldots, y_n) \cdot C = 0$.

$$\therefore \quad \sum_{i=1}^n m_j(y_i) C_i = 0$$

for any $m_j(x) \in \{m_0(x), \ldots, m_{M_q}(x)\}$

$$\therefore \quad \sum_{i=1}^n \varphi^j(y_i) C_i = 0 \quad \text{for any } j \in \{0, 1, \ldots, M_q\}$$

Therefore

$$\sum_{k,\ell=1}^n c_k c_\ell \tilde{K}(y_k, y_\ell) = \sum_{k,\ell=1}^n c_k c_\ell K(y_k, y_\ell) - \sum_{k,\ell=1}^n c_k c_\ell \varphi^i(y_k) K(y_\ell, a_i) - \sum_{k,\ell=1}^n c_k c_\ell \varphi^i(y_\ell) K(y_k, a_i) + \sum_{k,\ell=1}^n c_k c_\ell \varphi^i(y_k) \varphi^j(y_\ell) K(a_i, a_j)$$

where

$$\sum_{k,p=1}^{n} c_k c_p \varphi^i(y_k)\varphi^j(y_p) K(a_i,a_j)$$

$$= K(a_i,a_j)\underbrace{\left(\sum_{k=1}^{n} c_k \varphi^i(y_k)\right)}_{=0}\underbrace{\left(\sum_{p=1}^{n} c_p \varphi^j(y_p)\right)}_{=0} = 0$$

$$\sum_{k,p=1}^{n} c_k c_p \varphi^i(y_k) K(y_p, a_i)$$

$$= \sum_{p=1}^{n} c_p K(y_p,a_i)\underbrace{\left(\sum_{k=1}^{n} c_k \varphi^i(y_k)\right)}_{=0} = 0$$

and similarly $\sum_{k,p=1}^{n} c_k c_p \varphi^i(y_p) K(y_k, a_i) = 0.$

$$\therefore \sum_{k,p=1}^{n} c_k c_p \tilde{K}(y_k,y_p) = \sum_{k,p=1}^{n} c_k c_p K(y_k,y_p)$$

as was to be shown. ///

Remark:
What is interesting about $\sum_{k,p=1}^{n} c_k c_p \tilde{K}(y_k,y_p)$ is that it can be simplified further by dropping any term in $\tilde{K}(x,y)$ that has the form $\|x-y\|^{2j}$ where $j \le q$ ← the max monomial order.

This follows since

$$\|x-y\|^{2j} = \cdots + m(x)m'(y) + \cdots$$

where m and m' are monomials, one of which has has order $\le j$.

$$\therefore \sum_{k,p=1}^{n} c_k c_p \|y_k - y_p\|^{2j}$$

$$= \cdots + \underbrace{\left(\sum_{k=1}^{n} c_k m(y_k)\right)\left(\sum_{p=1}^{n} c_p m'(y_p)\right)}_{\text{either or is zero}} + \cdots \quad \text{(iii)}$$

$$= 0$$

This suggests that there exists other bivariate functions $\tilde{K}(x,y)$ which satisfy (ii) but which are not positive definite (except on increments that annihilate monomials of order $\le q$.)

A collection of generalized auto-cov funs

Suppose $K(t): \mathbb{R} \to \mathbb{R}$ is a symmetric function which has $2q$ finite derivatives at $t=0$.
Let

$$\mathcal{D}_q K(t) := K(t) - \sum_{k=0}^{q} \frac{K^{(2k)}(0)}{(2k)!} t^{2k}$$

Claim: If $K(t)$ is an auto-covariance function on $\mathbb{R}^d$ then $\mathcal{D}_q K(t)$ is a generalized auto-covariance function on $\mathbb{R}^d$ of order $m \ge q$

Proof: Let $x_1, \ldots, x_n \in \mathbb{R}^d$ and $c_1, \ldots, c_n \in \mathbb{R}$ such that

$$\sum_{i=1}^{n} c_i m(x_i) = 0$$

for all $m(\cdot) \in \{\text{monomials on } \mathbb{R}^d \text{ of order } \le q\}$.

By (iii) we have that

$$\underbrace{\sum_{k,p=1}^{n} c_k c_p K(\|x_k - x_p\|)}_{\ge 0 \text{ by positive definiteness}} = \underbrace{\sum_{k,p=1}^{n} c_k c_p \mathcal{D}_q K(\|x_k - x_p\|)}_{\ge 0} \Rightarrow$$

///

Example:

$$-\frac{1}{2}\overbrace{\gamma(\|x-y\|)}^{\text{variogram}} = K(\|x-y\|) - K(0)$$

$$= \mathcal{D}_q K(\|x-y\|) \text{ for } q=0$$

$\therefore$ $-\frac{1}{2}\gamma(\|x-y\|)$ is a generalized auto-covariance of order 0.

Example:

Consider the Matérn auto-cov $\mathcal{M}_\nu$.

By properties of $\mathcal{M}_\nu$

$|\mathcal{M}_\nu^{(2q)}(0)| < \infty$, $\forall$ integer $q < \nu$

$\Rightarrow$ $\mathcal{D}_{\lfloor\nu\rfloor}\mathcal{M}_\nu(\|x-y\|)$ is a generalized covariance function of order q for any $q \geq \lfloor\nu\rfloor$

Also recall from Lecture 1 that

$$\mathcal{M}_\nu(t) = \underbrace{\sum_{k=0}^{\lfloor\nu\rfloor} b_k t^{2k}}_{\text{order } 2\lfloor\nu\rfloor} + c_\nu \mathcal{G}_\nu(t) + \underbrace{\mathcal{O}(t^{2\lfloor\nu\rfloor+2})}_{\text{order } 2\lfloor\nu\rfloor+2}$$

as $t \to 0$ where $c_\nu > 0$ and

$$\mathcal{G}_\nu(t) := (-1)^{\lfloor\nu\rfloor+1} \times \begin{cases} t^{2\nu}\log t & \text{if } \nu = 1, 2, \ldots \\ t^{2\nu} & \text{otherwise} \end{cases}$$

These $\mathcal{G}_\nu$ are also generalized auto-covariance functions. To see why let $x_1, \ldots, x_n \in \mathbb{R}^d$ and $c_1, \ldots, c_n \in \mathbb{R}$ s.t.

$$\sum_{i=1}^n c_i\, m(x_i) = 0$$

for all $m(\cdot) \in \{\text{monomials on } \mathbb{R}^d \text{ of order} \leq \lfloor\nu\rfloor\}$.

Now

$$0 \leq \lim_{\varepsilon \downarrow 0} \sum_{i=1}^n c_i c_j \frac{\mathcal{M}_\nu(\varepsilon\|x_i - x_j\|)}{\varepsilon^{2\nu}}$$

$$= \lim_{\varepsilon \downarrow 0} \sum_{i=1}^n c_i c_j \frac{\mathcal{D}_{\lfloor\nu\rfloor}\mathcal{M}_\nu(\varepsilon\|x_i - x_j\|)}{\varepsilon^{2\nu}}$$

since $\sum_{i,j=1}^n c_i c_j \|x_i - x_j\|^{2k} = 0$ for any $k \leq \lfloor\nu\rfloor$

$$= \lim_{\varepsilon \downarrow 0} \sum_{i=1}^n c_i c_j \, c_\nu \frac{\mathcal{G}_\nu(\varepsilon\|x_i - x_j\|)}{\varepsilon^{2\nu}}$$

$$+ \lim_{\varepsilon \downarrow 0} \mathcal{O}\left(\frac{\varepsilon^{2\lfloor\nu\rfloor+2}}{\varepsilon^{2\nu}}\right)$$

Notice that $\lim_{\varepsilon \downarrow 0} \mathcal{O}\left(\frac{\varepsilon^{2\lfloor\nu\rfloor+2}}{\varepsilon^{2\nu}}\right) = 0$.

Also if $\nu \notin \{1, 2, \ldots\}$ then

$$\sum_{i,j=1}^n c_i c_j \frac{\mathcal{G}_\nu(\varepsilon\|x_i - x_j\|)}{\varepsilon^{2\nu}} = \sum_{i,j=1}^n c_i c_j (-1)^{\lfloor\nu\rfloor+1} \|x_i - x_j\|^{2\nu}$$

$$= \sum_{i,j=1}^n c_i c_j \mathcal{G}_\nu(\|x_i - x_j\|)$$

and if $\nu \in \{1, 2, \ldots\}$ then

$$\sum_{i,j=1}^n c_i c_j \frac{\mathcal{G}_\nu(\varepsilon\|x_i - x_j\|)}{\varepsilon^{2\nu}} = \sum_{i,j=1}^n c_i c_j (-1)^{\lfloor\nu\rfloor+1} \|x_i - x_j\|^{2\nu} \log\|x_i - x_j\|$$

$$+ \sum_{i,j=1}^n c_i c_j (-1)^{\lfloor\nu\rfloor+1} \|x_i - x_j\|^{2\nu} \log\varepsilon$$

This is zero since $2\nu = 2\lfloor\nu\rfloor$ and x_i, c_i annihilate polys of order $\lfloor\nu\rfloor$

$$= \sum_{i,j=1}^n c_i c_j \mathcal{G}_\nu(\|x_i - x_j\|)$$

Putting this all together we get

$$0 \le \lim_{\varepsilon \downarrow 0} \sum_{i,j=1} c_i c_j \frac{\mathcal{M}_\nu(\varepsilon \|x_i - x_j\|)}{\varepsilon^{2\nu}}$$

$$= \sum_{i,j=1}^{n} c_i c_j c_\nu \mathcal{Y}_\nu(\|x_i - x_j\|)$$

(c_ν: positive)

This establishes the following

Claim: For any $\nu > 0$ & $d \in \{1, 2, 3, \dots\}$

$$\mathcal{Y}_\nu(t) := (-1)^{\lfloor \nu \rfloor + 1} \times \begin{cases} t^{2\nu} \log t & \text{if } \nu = 1, 2, \dots \\ t^{2\nu} & \text{otherwise} \end{cases}$$

are generalized auto-covariance functions on $\mathbb{R}^d$ of order $q \ge \lfloor \nu \rfloor$.

Pre-Kriging REML parameter estimation

As is usually the case, the covariance function $K(x,y)$ used for Kriging has unknown parameters that need to be estimated from the data.

Write θ for a generic parameter vector and $K_\theta(x,y)$ for the dependence of $K(x,y)$ on θ.

Recall the regression characterization of the data $d_1, \dots, d_n$

$$d = \begin{bmatrix} d_1 \\ \vdots \\ d_n \end{bmatrix} \sim N\left(f(x_1, \dots x_n)^T \beta \,,\; \Sigma_\theta + \sigma_\varepsilon^2 I \right)$$

(Σ_θ: the matrix with $(i,j)^{th}$ entry $K_\theta(x_i, x_j)$)

If we construct a matrix M such that

$$M f(x_1, \dots, x_n)^T = \text{zero matrix}$$

then

$$Md \sim N\Big(\underbrace{M f(x_1, \dots, x_n)^T \beta}_{= \vec{0}}, M\left(\Sigma_\theta + \sigma_\varepsilon^2 I\right) M^T \Big)$$

so the likelihood of Md only depends on θ and not β. The REML (i.e. restricted max likelihood) estimate of θ is given by

$$\hat{\theta}_{REML} = \arg\max_\theta \log P(Md \mid \theta).$$

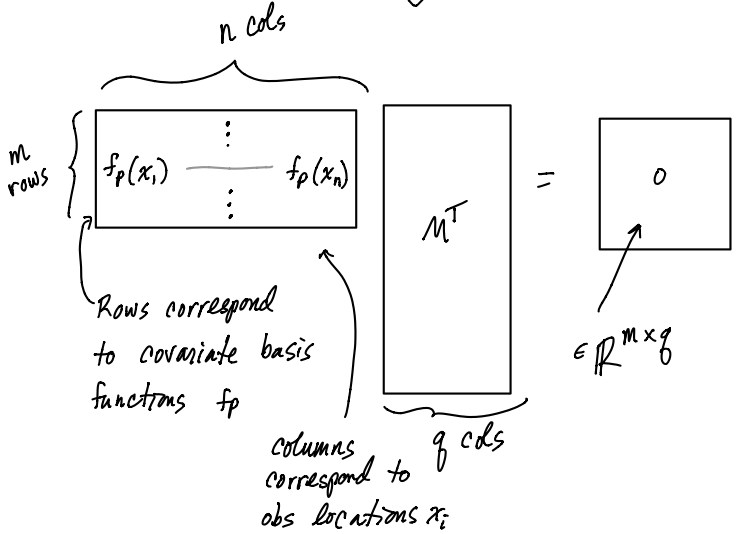
Notice Md will necessarily be a lossy compression. Lets see how many degrees of freedom we drop:

$$M\, f(x_1,\dots,x_n)^T = \text{zero matrix}$$

$$\Updownarrow$$

$$f(x_1,\dots,x_n)\, M^T = \text{zero matrix}$$

$$\Updownarrow$$



So the best we can hope for is an M where the columns of M^T spans the null space of $f(x_1,\dots,x_n)$ which has dimension at least $n-m$, i.e.

$$M\, f(x_1,\dots,x_n)^T = \text{zero matrix}$$

$$\Downarrow$$

$$\dim(Md) \geq n-m$$

Remark: A theme of the previous sections was that the final form of $\hat{Y}(x_0)$ or $\text{var}(\hat{Y}(x_0))$ doesn't require specifying everything about $K_\theta(x,y)$... just the parts that are not annihilated by linear combinations that satisfy

$$f(x_1,\dots,x_n)\cdot c = 0$$

or

$$f(x_1,\dots,x_n,x_0)\cdot \begin{bmatrix} \lambda \\ -1 \end{bmatrix} = 0$$

Notice that REML estimation is no different since we are using the likelihood of

$$Md \sim N\left(0,\, M\underbrace{\left(\Sigma_\theta + \sigma_\varepsilon^2 I\right)}M^T\right)$$

where $f(x_1,\dots,x_n)\, M^T = 0$

$$\Downarrow$$

$M\Sigma_\theta M^T$ annihilates parts of K_θ

REML + Kriging with $\mathcal{G}_\nu$ and parameter compression $\left(\frac{\sigma_Z^2}{\rho^{2\nu}}, \sigma_\varepsilon^2\right)$

Now we have a complete package:

- (cov) parameter estimation
- interpolation, $\widehat{Y(x_0)}$
- standard errors, $sd(\widehat{Y(x_0)})$

which is especially simplified using

- Generalized covariance Modeling with $K_\theta(x,y) \equiv \sigma^2 c_\nu \mathcal{G}_\nu\left(\sqrt{2\nu}\,\frac{\|x-y\|}{\rho}\right)$

$$\equiv \frac{\sigma^2}{\rho^{2\nu}} c_\nu (2\nu)^\nu \mathcal{G}(\|x-y\|)$$

when ν is an integer the increments annihilate the lower order term.

So the generalized covariance parameter vector is given by $\theta=(\theta_1,\theta_2)$

$$\theta_1 = \nu$$

$$\theta_2 = \frac{\sigma^2}{\rho^{2\nu}}$$

- $\{f_1(x),\dots,f_m(x)\}$ contain all monomials (in x) of order $\le \lfloor \nu \rfloor$

In this case there are then 3 parameters:

$$\theta_1 = \nu, \quad \theta_2 = \frac{\sigma^2}{\rho^{2\nu}}, \quad \sigma_\varepsilon^2 = \text{nugget variance}.$$

Warning about interpreting $\hat{\beta}$ when using generalized covariance functions with Kriging

Recall that Kriging is a method for estimating

$$Y(x) = \sum_{p=1}^{m} f_p(x)\beta_p + Z(x)$$

where $Z \sim GRF(0, K_\theta)$.

The fact that our goal is the same of the covariate term plus $Z(x)$, in essence, allows us to conflate the fixed effects:

$$\sum_{p=1}^{m} f_p(x)\beta_p$$

with the random effects

$$Z(x)$$

This means that $\hat{\beta}$ computed from K_θ will be different than $\hat{\beta}$ computed from a generalized cov G_θ, even though the predictions of $\hat{Y}$ yeild the same result.

Explicit form of the Principle irregular term for $M_\nu(t)$

Recall that

$$M_\nu(t) = \underbrace{\sum_{k=0}^{\lfloor \nu \rfloor} b_k t^{2k}}_{\text{order } 2\lfloor \nu \rfloor} + c_\nu \mathcal{Y}_\nu(t) + \underbrace{\mathcal{O}(t^{2\lfloor \nu \rfloor + 2})}_{\text{order } 2\lfloor \nu \rfloor + 2}$$

as $t \to 0$ where $c_\nu > 0$ and

$$\mathcal{Y}_\nu(t) := (-1)^{\lfloor \nu \rfloor + 1} \times \begin{cases} t^{2\nu} \log t & \text{if } \nu = 1, 2, \cdots \\ t^{2\nu} & \text{otherwise.} \end{cases}$$

It would be nice to know the constant term c_ν so one can compare $M_\nu(t)$ to its principle irregular term $c_\nu \mathcal{Y}_\nu(t)$.

The derivation is based on 9.6.2, 9.6.10 and 9.6.11 of Abramowitz and Stegun

Case 1: $\nu \notin \mathbb{Z}$

$$t^\nu K_\nu(t) = \frac{1}{2} \frac{\pi t^\nu}{\sin(\nu\pi)} \left(I_{-\nu}(z) - I_\nu(z) \right) \qquad (9.6.2)$$

$$= \frac{1}{2} \frac{\pi t^\nu}{\sin(\nu\pi)} \left(\left(\frac{t}{2}\right)^{-\nu} \sum_{k=0}^{\infty} \frac{\left(\frac{t}{2}\right)^{2k}}{k!\,\Gamma(-\nu+k+1)} - \left(\frac{t}{2}\right)^{\nu} \sum_{k=0}^{\infty} \frac{\left(\frac{t}{2}\right)^{2k}}{k!\,\Gamma(+\nu+k+1)} \right) \qquad (9.6.10)$$

$$= \frac{1}{2} \frac{\pi}{\sin(\nu\pi)} \left(2^\nu \sum_{k=0}^{\infty} \frac{\left(\frac{t}{2}\right)^{2k}}{k!\,\Gamma(-\nu+k+1)} \right)$$

$$+ \frac{1}{2} \frac{\pi}{\sin(\nu\pi)} \left(- \frac{t^{2\nu}}{2^\nu \Gamma(\nu+1)} \right) \quad \leftarrow \text{Principle irregular term (PIT)}$$

$$\frac{1}{2} \frac{\pi}{\sin(\nu\pi)} \left(- \frac{1}{2^\nu} \sum_{k=1}^{\infty} \frac{\left(\frac{t}{2}\right)^{2k+2\nu}}{k!\,\Gamma(+\nu+k+1)} \right)$$

$\therefore$

The PIT of $M_\nu(t)$

$$= \text{the PIT of } \frac{2^{1-\nu}}{\Gamma(\nu)} (\sqrt{2\nu}\, t)^\nu K_\nu(\sqrt{2\nu}\, t)$$

$$= \left(\frac{2^{1-\nu}}{\Gamma(\nu)} \right) \left(\frac{1}{2} \frac{\pi}{\sin(\nu\pi)} \right) \left(- \frac{(\sqrt{2\nu}\, t)^{2\nu}}{2^\nu \Gamma(\nu+1)} \right)$$

$$= \frac{\pi}{\sin(\nu\pi)} \left(\frac{-(\nu/2)^\nu}{\Gamma(\nu)\Gamma(\nu+1)} \right) t^{2\nu}$$

Case 2: $\nu \in \mathbb{Z}$

$$t^{\nu} K_{\nu}(t) = (\text{low order} + \text{higher order}) + t^{\nu}(-1)^{\nu+1} \log\left(\frac{t}{2}\right)\left(\frac{t}{2}\right)^{\nu} \frac{1}{\Gamma(\nu+1)}$$

by (9.6.11).

$\therefore$

The PIT of $M_{\nu}(t)$

= the PIT of $\frac{2^{1-\nu}}{\Gamma(\nu)}(\sqrt{2\nu}\,t)^{\nu} K_{\nu}(\sqrt{2\nu}\,t)$

$$= 2(-1)^{\nu}\left(\frac{-(\nu/2)^{\nu}}{\Gamma(\nu)\Gamma(\nu+1)}\right)\log(t)\,t^{2\nu}$$

In Summary

$$c_{\nu} Y_{\nu}(t) = \begin{cases} 2(-1)^{\nu} b_{\nu} \log(t)\, t^{2\nu}, & \nu \in \mathbb{Z} \\ \frac{\pi}{\sin(\nu\pi)} b_{\nu}\, t^{2\nu}, & \nu \notin \mathbb{Z} \end{cases}$$

where

$$b_{\nu} = \frac{-(\nu/2)^{\nu}}{\Gamma(\nu)\Gamma(\nu+1)}$$