FROM: Value of information in the Earth sciences, Eidsvik, Mukerji & Bhattacharjya, Cambridge, 2015

Discussion: variogram or covariance

Geostatisticians have traditionally preferred the variogram over the covariance/correlogram because (i) its inference does not call for prior inference of the mean and variance and (ii) there are theoretical models such as fractals that have infinite variance; hence, no covariance is defined, and yet they may have a finite variogram. However, in recent practice it is just as common to work with the covariance and estimate the mean as a function of explanatory variables as in a regression model.

Spatial two-point correlation between two different attributes y_1 and y_2 (e.g., porosity and permeability) is described by the cross-covariance function given by:

$$\hat{C}_{12}(\boldsymbol{t}) = \frac{1}{n(\boldsymbol{t})}\sum_i \big(y_1(\boldsymbol{s}_i+\boldsymbol{t})-\mu_1\big)\big(y_2(\boldsymbol{s}_i)-\mu_2\big), \tag{4.6}$$

where μ_1 and μ_2 are the stationary means of the two attributes.

Unless data locations are on a regular grid, tolerance of lag value (and direction) is needed to find enough pairs, $n(\boldsymbol{t})$, of data approximately at distance $\boldsymbol{t}$ apart to infer an experimental value for the variogram in Equation (4.3), covariance in Equation (4.4), and cross-covariance in Equation (4.6). The level of tolerance needed depends on the amount and spatial layout of the available data. The tolerance setting is another modeling choice that impacts the resulting experimental variograms and hence the analytical models used to fit the experimental variogram. A full angle/direction tolerance pools together all data pairs with the same distance modulus $|\boldsymbol{t}|$ irrespective of direction, and the result is then an omni-directional experimental variogram often modeled as an isotropic model. **Isotropy** – that is, invariance with direction – is then a consequence of the angle tolerance, not necessarily a physical characteristic of the underlying phenomenon.

In practice, when we have data $y(\boldsymbol{s}_1), y(\boldsymbol{s}_2),..,y(\boldsymbol{s}_n)$ available, the empirical variogram is constructed. Next, we fit a parametric model, such as those presented in Table 4.1 (or any other legitimate model) to the data. The simplest way to do this is by visual inspection to answer questions such as: does the decline with distance appear to follow an exponential trend? And is there a nugget effect? More sophisticated ways of parameter estimation will be treated in Section 4.4 in the context of Gaussian random fields.

Four common examples of spatial covariance functions are shown in Table 4.1. Many other popular spatial covariance functions are described in standard books on geostatistics – e.g., Deutsch and Journel (1992), Goovaerts, (1997), Lantuejoul, (2002), and Chiles and Delfiner (2012). In the equations in Table 4.1, τ^2 is the nugget effect, which only affects the variance; σ^2 is the variance–covariance of the spatially dependent process, so the overall variance equals $\tau^2+\sigma^2$; and η determines the decay of the covariance function. If η is large, the covariance goes quickly to 0, while it decays more slowly for small η. For the exponential covariance function, one can parametrize the decay by the effective spatial range $3/\eta$, since $\exp(-3) \approx 0.05$, indicating that the correlation is only 0.05 at spatial distance $|\boldsymbol{t}| = 3/\eta$. The four covariance functions are displayed as a function of lag distance in Figure 4.3 (left) together with the associated variograms (Figure 4.3, right). The model parameters were here set to $\tau^2 = 0.5^2$, $\sigma^2 = 1^2$, and $\eta = 10$. All these four common parametric covariance functions give valid positive definite covariance matrices for n response variables at any combination of spatial locations $\boldsymbol{s}_1,...,\boldsymbol{s}_n$.

The connection between the covariance models and the related variogram model is shown in Figure 4.3. The variogram model $\gamma(|\boldsymbol{t}|) = C(0) - C(|\boldsymbol{t}|)$ would approach

Table 4.1. *Four examples of spatial covariance functions. For all covariance models, the first term represents a nugget effect tied to uncorrelated noise or measurement noise. The latter terms include spatial correlation*

Model	Covariance
Exponential	$C(\lvert\boldsymbol{t}\rvert) = \tau^2 I(\lvert\boldsymbol{t}\rvert=0) + \sigma^2 \exp(-\eta\lvert\boldsymbol{t}\rvert)$
Matern 3/2	$C(\lvert\boldsymbol{t}\rvert) = \tau^2 I(\lvert\boldsymbol{t}\rvert=0) + \sigma^2 (1+\eta\lvert\boldsymbol{t}\rvert)\exp(-\eta\lvert\boldsymbol{t}\rvert)$
Cauchy type	$C(\lvert\boldsymbol{t}\rvert) = \tau^2 I(\lvert\boldsymbol{t}\rvert=0) + \sigma^2 \frac{1}{(1+\eta\lvert\boldsymbol{t}\rvert)^3}$
Gaussian	$C(\lvert\boldsymbol{t}\rvert) = \tau^2 I(\lvert\boldsymbol{t}\rvert=0) + \sigma^2 \exp(-\eta^2\lvert\boldsymbol{t}\rvert^2)$

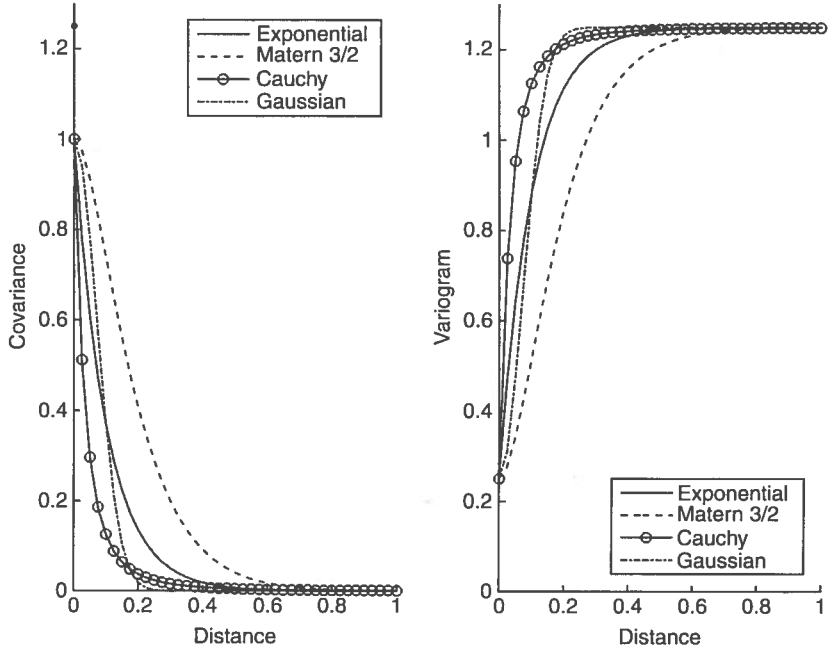

Figure 4.3 Three different covariance functions and variograms. Left: covariance functions plotted as a function of distance (first axis). Right: variogram plotted as a function of distance (first axis). The parameters are variance $\sigma^2 = 1$ for the random effect, $\tau^2 = 0.5^2$ for the nugget effect, and correlation decay $\eta = 10$.

and as more stochastic modeling techniques are developed, the most successful case studies will be those that view the assortment of spatial modeling methods as a tool kit rather than as a silver bullet.

4.4 Gaussian models

The Gaussian random field model is a key construction in geostatistics and is often used as a self-standing model, especially for very-high-dimensional models or very large data sizes. Moreover, it is common to apply the Gaussian model as a building block when constructing hierarchical model formulations. In this section, we discuss traditional but important aspects of the Gaussian model in spatial statistics.

4.4.1 The spatial regression model

The spatial regression model for continuous response data relies on Gaussian variables and linear association. This is possibly the most commonly applied model in spatial statistics – see, e.g., Cressie (1993), Stein (1999), and Banerjee et al. (2004). The model has historical importance related to (universal) Kriging. Moreover, the model provides a natural extension of the usual linear regression model, as it also accounts for the spatially correlated error terms. The attractive computational properties of the Gaussian model make it one of the few applicable models for massive data sets. The Gaussian assumption can often be justified by the central limit theorem, stating that sums and means of random variables converge to Gaussian variables. Thus, even though the response is not really Gaussian, these assumptions can provide useful results in many situations.

The response variable is assumed to be partially explained by (i) explanatory variables, (ii) a smooth Gaussian noise process – a Gaussian random field, and (iii) independent errors. The idea is to incorporate the spatial smoothness and, as a result, obtain more reliable estimates of the regression parameters for improved predictions. Ignoring a spatially structured noise term could give biased estimates and erroneous uncertainty bounds. The following exposition is based on a univariate response variable. At the end of the section, a more general framework is presented. We assume that the process is defined at all locations $s \in \mathcal{D}$, where $\mathcal{D}$ denotes a continuous spatial domain in two or three dimensions. The model for the response $y(s)$ at an arbitrary site is

$$y(s) = h^t(s)\beta + w(s) + \varepsilon(s), \tag{4.10}$$

where $\beta = (\beta_1, \ldots, \beta_k)$ is the vector of k regression parameters and $h(s) = (h_1(s), \ldots, h_k(s))$ is the vector of k covariates at site s. The residual is split into two parts: $w(s)$ and $\varepsilon(s)$. The spatially structured residual $w(s)$ provides dependence, $\mathrm{Cov}(w(s), w(s')) = \Sigma(s, s')$, capturing the effect of unobserved covariates with a spatial pattern. Statisticians often refer to the regression parameters β as fixed effects, while the structured Gaussian field $w(s)$ is the random effects. The non-structured spatial residual $\varepsilon(s)$ is independent white noise with

$\mathrm{Var}(\varepsilon(s)) = \tau^2$, which can be interpreted as the measurement error. The spatial regression model in Equation (4.10) can also be written as:

$$y(s) = x(s) + N(0, \tau^2), \quad E(x(s)) = h^t(s)\beta, \quad \mathrm{Cov}(x(s), x(s')) = \Sigma(s, s'). \tag{4.11}$$

The latent process is often of key interest to the decision maker. In Equation (4.11), this latent process $x(s)$ is imperfectly observed by $y(s)$.

Discussion: spatial model versus ordinary least squares

The basic linear regression model, which forms the starting point in many contexts of exploratory data analysis, assumes independent non-structured error terms. Here, in the spatial context, the noise process is modeled using the tools of variograms and covariance functions described previously. The covariance structure of the spatial residual $w(s)$ is thus typically characterized by a few parameters describing the scale and correlation range. Customarily, the spatial covariance $\mathrm{Cov}(w(s), w(s')) = \Sigma(s, s')$ is modeled by a stationary, isotropic process – i.e., it only depends on the absolute distance between the locations s and s' (see Table 4.1). For the independent noise process, we assume that $\varepsilon(s) \sim N(0, \tau^2)$ for all s. This was interpreted as the nugget effect in Section 4.2.

Assume that we can observe the spatial process $y(s)$ and associated covariates $h^t(s)$ at n locations $s_1, \ldots, s_n$. Under the specified assumptions, we can now write the Gaussian regression model as a **hierarchical model**. Let us denote the collection of data by length n vector $y = (y(s_1), \ldots, y(s_n))$; the latent random effects by $x = (x(s_1), \ldots, x(s_n))$; and the covariates by a size $n \times k$ matrix H, where row i is $h^t(s_i)$. Then

$$p(x) = N(H\beta, \Sigma), \qquad p(y \mid x) = N(x, \tau^2 I_n), \tag{4.12}$$

and the marginal pdf of the response (integrating out the random effects) becomes

$$p(y) = N(H\beta, C), \qquad C = C(\theta) = \Sigma + \tau^2 I_n, \tag{4.13}$$

where C is a $n \times n$ covariance matrix, and $\theta = (\sigma^2, \eta, \tau^2)$ denotes the set of covariance parameters.

The set of observation locations is called the spatial design. A regular design has observation sites $s_1, \ldots, s_n$ on a grid of the spatial domain $\mathcal{D}$. Irregular sampling designs are, however, more common in practice. For instance, in a situation with monitoring sites for precipitation and wind or air pollution, one would tend to place the sites near roads or cities for logistical reasons. On the other hand, remotely sensed data such as satellite data or seismic data are often processed to be represented on a regular grid.

Figure 4.4 shows a realization of a Gaussian random field on the unit square. This is shown on a regular grid in the left display. An irregular sampling of the data of a much

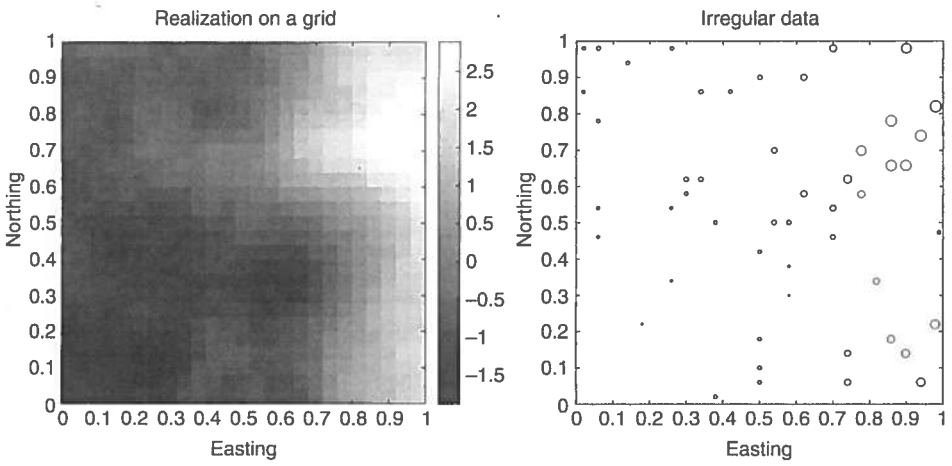

Figure 4.4 Realization of a Gaussian process in two spatial dimensions. Left: the spatial variable represented on a dense regular grid. Right: the spatial variable is represented at only 100 irregular locations.

smaller size is also illustrated in the right display. In Figure 4.4, the model is based on using the east and north coordinates as covariates, and the regression parameters are set to $\beta = (-2,3,1)$, where the first entry corresponds to an intercept term at the origin and the next two are the spatial covariates. The covariance function is a Matern type with smoothness parameter (3/2), as in Table 4.1, and with parameters $\sigma^2 = 0.5^2$, $\eta = 9$, and $\tau^2 = 0.05^2$, which corresponds to a correlation range of about a third of the unit square. Note that there is only a small nugget effect here. In Figure 4.4, we see that the random field increases with the east (and north) coordinate with smooth variability defined by the Gaussian residual process.

We assume that the parameters β and θ are fixed but unknown. These parameters must be estimated based on the data and explanatory variables. A common way of specifying the parameter values is by maximum likelihood estimation (MLE), described in Section 2.4 and Appendix A.1. The Gaussian distribution defines the log-likelihood as a function of parameters β and θ – i.e.,

$$l(\theta,\beta) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log|C(\theta)| - \frac{1}{2}(y - H\beta)' C^{-1}(\theta)(y - H\beta). \quad (4.14)$$

The maximum likelihood estimates are defined by

$$(\hat{\beta},\hat{\theta}) = \operatorname{argmax} l(\theta,\beta). \quad (4.15)$$

An algorithm for locating the maximum is presented in Appendix A.1. For fixed covariance parameters θ, the estimate for the regression parameters is

$$\hat{\beta} = \hat{\beta}(y;\theta) = (H'C^{-1}H)^{-1} H'C^{-1}y, \qquad \operatorname{Var}(\hat{\beta}) = (H'C^{-1}H)^{-1}. \quad (4.16)$$

4.4.2 Optimal spatial prediction: Kriging

We showed how Kriging is the method for optimal pointwise spatial prediction in the Gaussian model. Suppose that we wish to predict responses $y_0 = (y(s_{0,1}),\ldots,y(s_{0,N}))$ at N prediction sites $s_{0,j}$, $j = 1,\ldots,N$, given data $y = (y(s_1),\ldots,y(s_n))$ at n observation sites. The joint distribution of y_0 and y is Gaussian with the following mean and covariance:

$$p(y_0, y) = N\left(\begin{pmatrix} H_0 \\ H \end{pmatrix}\beta, \begin{pmatrix} C_0 & C_{0,*} \\ C_{0,*}' & C \end{pmatrix}\right). \quad (4.17)$$

The $N \times k$ matrix H_0 contains the covariates at the prediction sites. Moreover, the $N \times N$ matrix C_0 is the covariance matrix for the responses at all prediction sites, while the $N \times n$ matrix $C_{0,*}$ contains the covariances between the N variables at prediction sites and the n variables at observation sites. The covariance matrices depend on the statistical model parameters θ, but the cross-covariance matrix $C_{0,*}$ does not depend on the measurement error variance τ^2 since we assume independent nugget effects.

Recall from Chapter 2 that the conditional pdf of y_0 given y (and for fixed β and θ) is also Gaussian. The length N vector of conditional means or **Kriging predictions** is

$$E(y_0 \mid y) = H_0\beta + C_{0,*} C^{-1}(y - H\beta), \quad (4.18)$$

and the associated $N \times N$ conditional covariance is

$$\operatorname{Var}(y_0 \mid y) = C_0 - C_{0,*} C^{-1} C_{0,*}'. \quad (4.19)$$

The conditional variances are defined by the diagonal elements of this matrix. See Appendix A.1 for further details about these properties of the Gaussian pdf.

Discussion: interpreting the Kriging prediction variance

For sites that are close to other data, there is high correlation in $C_{0,*}$, and conditioning will reduce the variances in C_0 substantially. Sites that are farther from the observation sites will have larger prediction variances. The reduction of prediction variance also depends on the clustering of the observation sites according to C^{-1}. Two data at almost the same location will not contribute twice the information, since the two observations will be correlated and therefore somewhat redundant. It is quite remarkable that the Kriging variances do not depend on the data – they only depend on the geographic locations of the data and the prediction site. This holds for the Gaussian situation but may not hold in non-Gaussian settings where a large (or small) observation may influence the prediction variance as well.

Based on the modeling assumptions, the prediction distribution is also Gaussian. The 5th and 95th percentiles of the standard Gaussian distribution are −1.64 and 1.64. A 90% prediction interval for the response $y(s_{0,j})$ at the prediction site $s_{0,j}$ is then

$$\left(E(y(s_{0,j}) \mid y) - 1.64\sqrt{\operatorname{Var}(y(s_{0,j}) \mid y)},\ E(y(s_{0,j}) \mid y) + 1.64\sqrt{\operatorname{Var}(y(s_{0,j}) \mid y)}\right). \quad (4.20)$$

mean gets extremely small in high dimensions as it is always in the tail of the distribution. The probability of all independent variables occurring within the unit square is also displayed in Table A.1.

Consider also the linear transformation $y = Fx + e$, where $p(e) = N(0, T)$ and x and e are independent. The pdf of y is $p(y) = N(F\mu, F\Sigma F^T + T)$. The joint distribution of x and y is Gaussian with mean $\begin{pmatrix} \mu \\ F\mu \end{pmatrix}$ and covariance $\begin{pmatrix} \Sigma & \Sigma F^T \\ F\Sigma & F\Sigma F^T + T \end{pmatrix}$. The conditional distribution of x given y is also Gaussian with mean and variance:

$$\begin{aligned} \mu_{x|y} &= \mu + \Sigma F^T \left(F\Sigma F^T + T\right)^{-1} (y - F\mu), \\ \Sigma_{x|y} &= \Sigma - \Sigma F^T \left(F\Sigma F^T + T\right)^{-1} F\Sigma. \end{aligned} \tag{A.10}$$

Parameter estimation

For multiple data sets of $x^1, \ldots, x^B$, the empirical mean and covariance estimates are

$$\hat{\mu} = \frac{1}{B}\sum_{b=1}^{B} x^b, \quad \hat{\Sigma} = \frac{1}{B}\sum_{b=1}^{B} (x^b - \hat{\mu})(x^b - \hat{\mu})^t. \tag{A.11}$$

Suppose instead that the mean is modeled by $\mu = H\beta$, where the size $n \times k$ matrix H consists of known explanatory variables and that β is a length k vector of fixed but unknown regression parameters. If the data contain total perfect information $y = x$ and the prior covariance Σ is known, the maximum likelihood estimator (MLE) of the regression parameter β equals that of the weighted least-squares estimate:

$$\tilde{\beta} = \left(H^t \Sigma^{-1} H\right)^{-1} H^t \Sigma^{-1} x. \tag{A.12}$$

If the covariance matrix is $\Sigma = \sigma^2 I$, this simplifies to the ordinary least-squares $\tilde{\beta} = \left(H^t H\right)^{-1} H^t x$. An estimate of the noise level is $\hat{\sigma}^2 = \frac{1}{n}\left(x - H\tilde{\beta}\right)^t \left(x - H\tilde{\beta}\right)$. This estimate is asymptotically unbiased, but for a finite sample we may use $n - k$ in the denominator instead to achieve an unbiased variance estimate.

Assume that we have imperfect information $y = (y_1, \ldots, y_m)$ and the model is

$$p(x) = N(H\beta, \Sigma), \qquad p(y \mid x) = N(Fx, T). \tag{A.13}$$

The marginal likelihood of the data is $p(y) = N(G\beta, C)$ for $G = FH$ and $C = F\Sigma F^T + T$. The log-likelihood as a function of β and unknown fixed nuisance parameters θ in the prior covariance matrix $\Sigma = \Sigma(\theta)$, and/or the likelihood noise matrix $T = T(\theta)$, becomes

$$l(\theta, \beta) = -\frac{m}{2}\log(2\pi) - \frac{1}{2}\log|\mathbf{C}| - \frac{1}{2}(y - G\beta)^t \mathbf{C}^{-1}(y - G\beta). \tag{A.14}$$

The MLEs of β and θ are obtained by

$$\left(\hat{\beta}, \hat{\theta}\right) = \text{argmax}\ \left\{l(\theta, \beta)\right\}. \tag{A.15}$$

For fixed θ, the MLE of β is analytically available. We have

$$\frac{dl}{d\beta} = G^t \mathbf{C}^{-1} y - G^t \mathbf{C}^{-1} G\beta = \mathbf{0}, \quad \hat{\beta} = \hat{\beta}(y; \theta) = \left(G^t \mathbf{C}^{-1} G\right)^{-1} G^t \mathbf{C}^{-1} y. \tag{A.16}$$

A direct calculation shows that $Var(\hat{\beta}) = \left(G^t \mathbf{C}^{-1} G\right)^{-1}$.

Treating the regression parameter β as fixed, the MLE of nuisance parameters θ can be obtained by numerical maximization. Set $z = y - G\beta$, and let $\mathbf{Q} = \mathbf{C}^{-1}$. Denote a component of θ by θ_r, $r = 1, \ldots, d$. (In our setting of Chapter 4, the components are prior variance σ^2, correlation decay η, nugget τ^2, and $d = 3$.) The first derivative (score) of the log-likelihood in Equation (A.14) with respect to element θ_r becomes

$$\frac{dl}{d\theta_r} = -\frac{1}{2}\text{trace}\left(\mathbf{Q}\frac{d\mathbf{C}}{d\theta_r}\right) + \frac{1}{2} z^t \mathbf{Q}\frac{d\mathbf{C}}{d\theta_r}\mathbf{Q}z, \tag{A.17}$$

where we have used $\frac{d \log|\mathbf{C}|}{d\theta_r} = \text{trace}\left(\mathbf{Q}\frac{d\mathbf{C}}{d\theta_r}\right)$, $\frac{dz^t \mathbf{C}^{-1} z}{d\theta_r} = -z^t\left(\mathbf{Q}\frac{d\mathbf{C}}{d\theta_r}\mathbf{Q}\right)z$.

The expected Hessian is

$$E\left(\frac{d^2 l}{d\theta_r d\theta_{r'}}\right) = -\frac{1}{2}\text{trace}\left(\mathbf{Q}\frac{d\mathbf{C}}{d\theta_{r'}}\mathbf{Q}\frac{d\mathbf{C}}{d\theta_r}\right), \tag{A.18}$$

where we used $E(z^t \Sigma z) = \text{trace}\left(\Sigma Var(z)\right)$, assuming $E(z^t) = \mathbf{0}$.

Algorithm: the iterative Fisher scoring algorithm for obtaining the MLE of β and θ

Initiate β^0, θ^0 by least-squares estimation and empirical variograms or other approaches.
Iterate for $b = 0, 1, \ldots$, until convergence:

$\mathbf{C} = \mathbf{C}(\theta^b)$

$\beta^{b+1} = \beta^{b+1}(y; \theta) = [G^t \mathbf{C}^{-1} G]^{-1} G^t \mathbf{C}^{-1} y$,

$z = y - G\beta^{b+1}$

$\mathbf{Q} = \mathbf{C}^{-1}, \mathbf{C}_r^* = \frac{d\mathbf{C}(\theta^b)}{d\theta_r}, r = 1, \ldots, d$

$u_r = \frac{dl}{d\theta_r} = -\frac{1}{2}\text{trace}\left(\mathbf{Q}\mathbf{C}_r^*\right) + \frac{1}{2} z^t \mathbf{Q}\mathbf{C}_r^* \mathbf{Q}z, \qquad r = 1, \ldots, d$

$V_{\{rr'\}} = E\left(\frac{d^2 l}{d\theta_r d\theta_{r'}}\right) = -\frac{1}{2}\text{trace}\left(\mathbf{Q}\mathbf{C}_r^*\mathbf{Q}\mathbf{C}_{r'}^*\right), \qquad r, r' = 1, \ldots, d.$

$\theta^{b+1} = \theta^b + \mathbf{V}^{-1}\mathbf{u}.$

$b = b + 1.$