We have that the estimated density from the kernel density estimator is given by

$$\hat{f}_\theta(x) = \frac{1}{n}\sum_{i=1}^{n} K_\theta(x - x_i') \tag{1}$$

where K_h is the kernel function, and θ are the parameters, and x_i' is the ith training point. In our case x_i' is a m dimensional vector representing the daily change for each of the m stocks in our dataset. We assume that that each kernel function is normalized $\int_{\mathbb{R}^d} K_h(x)dx = 1$

Optimizing the Kernel Parameters

We would want to minimize the weighted log likelihood function, ie we would want to minimize:

$$\mathcal{L}(x_1', \dots, x_k') = -\sum_{i=1}^{k} w_i \log(\hat{f}_\theta(x_i')) \tag{2}$$

Where are w_i are weights of the ith training point, we will experiments with different weighting mechanisms later on.

Gaussian Kernel

Let us consider the case of a gaussian kernel

$$K_\Sigma(x) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac{1}{2}x^T\Sigma^{-1}x\right)$$

Where Σ is the covariance matrix for the kernel. Because the covariance matrix is symmetric positive semidefinite we can express Σ as $\Sigma = R^T R$ through Cholesky decomposition. We have that the derivative of the weighted log likelihood function is given by:

$$\frac{\partial}{\partial R}\mathcal{L}(x_1', \dots, x_k') = -\sum_{i=1}^{k} w_i \frac{1}{\hat{f}_\theta(x_i')}\frac{\partial \hat{f}_\theta(x_i')}{\partial R}$$

We have that

$$\begin{aligned}\frac{\partial|\Sigma|}{\partial R} &= \frac{\partial|R^T R|}{\partial R} \\ &= 2|\Sigma|R^{-T}\end{aligned}$$

Therefore we have that

$$\frac{\partial}{\partial R}\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} = -R^{-T}\frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \tag{3}$$

We also have that:

$$\frac{\partial}{\partial R}e^{\frac{1}{2}x^T\Sigma^{-1}x} = \frac{1}{2}e^{\frac{1}{2}x^T\Sigma^{-1}x}\frac{\partial}{\partial R}x^T(R^{-1}R^{-T})x \tag{4}$$

$\frac{\partial}{\partial R}x^T(R^{-1}R^{-T})x$ is very difficult to calculate, so we must approximate it. First we note that for a function $f(\mathbf{x})$ that takes in a vector $\mathbf{x}$ we have that the first order taylor expansion is given by:

$$f(\mathbf{x} + \Delta\mathbf{x}) \approx f(\mathbf{x}) + \frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}\Delta\mathbf{x} \tag{5}$$

Where $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$ is a $1 \times d$ vector, if $\mathbf{x}$ is a d dimensional vector. Therefore we argue that a generalization to a function of a matrix $\mathbf{X}$ is given by:

$$f(\mathbf{X} + \Delta\mathbf{X}) \approx f(\mathbf{X}) + \mathbf{1}^T\left(\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \circ \Delta\mathbf{X}\right)\mathbf{1} \tag{6}$$

Where $\circ$ is the Hadamard product, and $\mathbf{1}$ is a $d \times 1$ vector of ones. We note that $\mathbf{1}^T \left(\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}} \circ \Delta \mathbf{X}\right) \mathbf{1}$ equals to $\operatorname{tr}\left(\left(\frac{\partial f(\mathbf{X})}{\partial \mathbf{X}}\right)^T \mathbf{X}\right)$. We have for our specific case:

$$x^T((R+\delta R)^{-1}(R+\delta R)^{-T})x \approx x^T(R^{-1}R^{-T})x + \operatorname{tr}\left(\left(\frac{\partial(x^T R^{-1}R^{-T})x}{\partial R}\right)^T \delta R\right)$$

We note that for small pertubations, $(R+\delta R)^{-1} \approx R^{-1} - R^{-1}\delta R R^{-1}$, and therefore:

$$x^T(R^{-1} - R^{-1}\delta R R^{-1})(R^{-T} - R^{-T}\delta R^T R^{-T})x \approx x^T(R^{-1}R^{-T})x + \operatorname{tr}\left(\left(\frac{\partial(x^T R^{-1}R^{-T})x}{\partial R}\right)^T \delta R\right)$$

Only keeping the zeroth order and first order terms, we have that:

$$\begin{aligned} x^T(R^{-1}R^{-T})x - x^T\left(R^{-1}\delta R R^{-1}R^{-T} + R^{-1}R^{-T}\delta R^T R^{-T}\right)x \approx & x^T(R^{-1}R^{-T})x \\ & + \operatorname{tr}\left(\left(\frac{\partial(x^T R^{-1}R^{-T})x}{\partial R}\right)^T \delta R\right) \end{aligned}$$

$$-x^T\left(R^{-1}\delta R R^{-1}R^{-T} + R^{-1}R^{-T}\delta R^T R^{-T}\right)x \approx \operatorname{tr}\left(\left(\frac{\partial(x^T R^{-1}R^{-T})x}{\partial R}\right)^T \delta R\right) \tag{7}$$

Because the left side is a scalar, we can apply an trace operator to both sides, and noting that $R^{-1}R^{-T} = \Sigma^{-1}$, we have that:

$$\begin{aligned} -\operatorname{tr}\left(x^T\left(R^{-1}\delta R\Sigma^{-1} + \Sigma^{-1}\delta R^T R^{-T}\right)x\right) &\approx \operatorname{tr}\left(\left(\frac{\partial(x^T R^{-1}R^{-T})x}{\partial R}\right)^T \delta R\right) \\ -\operatorname{tr}\left(x^T R^{-1}\delta R\Sigma^{-1}x\right) - \operatorname{tr}\left(x^T\Sigma^{-1}\delta R^T R^{-T}x\right) &\approx \\ -\operatorname{tr}\left(x^T R^{-1}\delta R\Sigma^{-1}x\right) - \operatorname{tr}\left(x^T R^{-1}\delta R\Sigma^{-T}x\right) &\approx \\ -2\operatorname{tr}\left(x^T R^{-1}\delta R\Sigma^{-1}x\right) &\approx \\ -2\operatorname{tr}\left(\Sigma^{-1}xx^T R^{-1}\delta R\right) &\approx \operatorname{tr}\left(\left(\frac{\partial(x^T R^{-1}R^{-T})x}{\partial R}\right)^T \delta R\right) \end{aligned}$$

Therefore we can see that

$$\frac{\partial}{\partial R}x^T(R^{-1}R^{-T})x \approx -2\Sigma^{-1}xx^T R^{-1} \tag{8}$$

Therefore we have that:

$$\frac{\partial}{\partial R}e^{\frac{1}{2}x^T\Sigma^{-1}x} \approx -e^{\frac{1}{2}x^T\Sigma^{-1}x}\Sigma^{-1}xx^T R^{-1} \tag{9}$$

Therefore we have that

$$\frac{\partial}{\partial R}K_\Sigma(x) = -K_\Sigma(x)R^{-T} - K_\Sigma(x)\Sigma^{-1}xx^T R^{-1}$$