

GP Write Up 8

October 24, 2018

Loss function based on likelihood and Smoothing

1. Loss function based on likelihood

We defined a new loss function based on the joint distribution:

For a set of data points X and covariance matrix K , with $y = \mathbf{f} + \epsilon$ where $\mathbf{f} \sim \mathcal{N}(0, K(X, X))$ and $\epsilon \sim \mathcal{N}(0, \sigma_n^2)$ we can write

$$\log P(\mathbf{f}|X) \propto \frac{-1}{2} \mathbf{f}^T K^{-1} \mathbf{f} \quad (1)$$

$$\log P(y|\mathbf{f}, X) \propto \frac{-1}{2\sigma_n^2} (y - \mathbf{f})^T (y - \mathbf{f}) \quad (2)$$

$$\log P(\mathbf{f}, y|X) \propto \frac{-1}{2} \mathbf{f}^T K^{-1} \mathbf{f} - \frac{1}{2\sigma_n^2} (y - \mathbf{f})^T (y - \mathbf{f}) \quad (3)$$

we introduce a new variable \mathbf{v} such that $K\mathbf{v} = \mathbf{f}$ and rewrite equation (4) as :

$$\log P(\mathbf{v}, y|X) \propto \frac{-1}{2} \mathbf{v}^T K \mathbf{v} - \frac{1}{2\sigma_n^2} (y - K\mathbf{v})^T (y - K\mathbf{v}) \quad (4)$$

the optimization problem w.r.t to equation (4) is:

$$\theta_* = \operatorname{argmin}_{\theta} \frac{-1}{2} \mathbf{f}^T K_{\theta}^{-1} \mathbf{f} - \frac{1}{2\sigma_n^2} (y - \mathbf{f})^T (y - \mathbf{f}) \quad (5)$$

where θ has the parameters for the kernel function and the optimization problem w.r.t to (5) is:

$$\mathbf{v}_* = \operatorname{argmin}_{\mathbf{v}} \frac{-1}{2} \mathbf{v}^T K_{\theta} \mathbf{v} - \frac{1}{2\sigma_n^2} (y - K_{\theta} \mathbf{v})^T (y - K_{\theta} \mathbf{v}) \quad (6)$$

2. Spectral decomposition of K

The covariance matrix K can be decomposed as follows:

$$K = \mathbf{U} \lambda \mathbf{U}^T = \mathbf{U} \lambda \mathbf{U}^{-1} \quad (7)$$

where λ contains the eigenvalues and \mathbf{U} contains the eigenvectors. The columns of \mathbf{U} are orthonormal and hence $\mathbf{U}^T = \mathbf{U}^{-1}$.

Let $\mathbf{x}' = K\mathbf{x} = \mathbf{U} \lambda \mathbf{U}^{-1} \mathbf{x}$.

So the vector \mathbf{x} first gets transformed into a vector in the space where the columns are the basis vectors. This vector is then scaled in every dimension with the corresponding eigenvalue. The initial transformation is undone with the final \mathbf{U} .

We can write:

$$\mathbf{x}^T K \mathbf{x} = \mathbf{x}^T \mathbf{x}' = \mathbf{x} \cdot \mathbf{x}' \quad (8)$$

If we have a covariance function $k_\theta(\cdot, \cdot)$ in our initial formulation of the problem:

$$\operatorname{argmin}_\theta \mathbf{f}^T K_\theta^{-1} \mathbf{f} = \operatorname{argmin}_\theta \mathbf{f} \cdot \mathbf{f}' \quad (9)$$

We are trying to find \mathbf{f}' that minimizes (9) by increasing the angle between \mathbf{f} and \mathbf{f}' or by decreasing the magnitude of \mathbf{f}' or both.

In our formulation of the optimization problem:

$$\operatorname{argmin}_v \mathbf{v}^T K \mathbf{v} = \operatorname{argmin}_v \mathbf{v} \cdot \mathbf{v}' \quad (10)$$

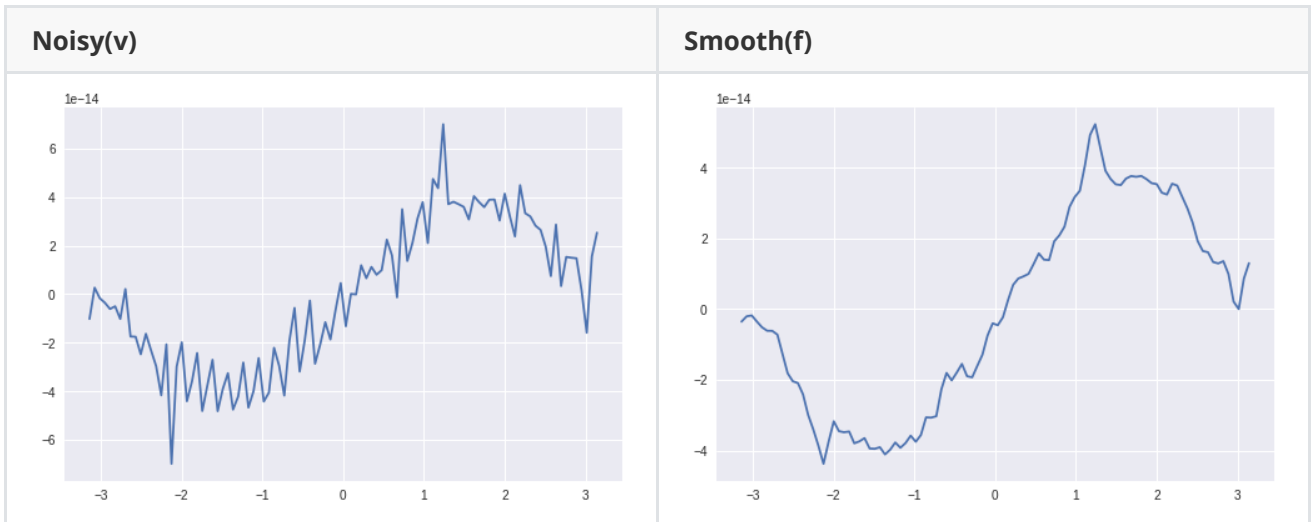
We are trying to find \mathbf{v} that decreases the dot product between \mathbf{v} and \mathbf{v}' . There can be infinite number of vectors \mathbf{v} with the trivial solution being a null vector but that is prevented by the regularization term $\frac{1}{2\sigma_n^2}(\mathbf{y} - K_\theta \mathbf{v})^T (\mathbf{y} - K_\theta \mathbf{v})$.

Initially we assumed that $\mathbf{v} = K^{-1} \mathbf{f}$, so the optimization problem minimizes \mathbf{v} to \mathbf{f}' and \mathbf{v}' to \mathbf{f} .

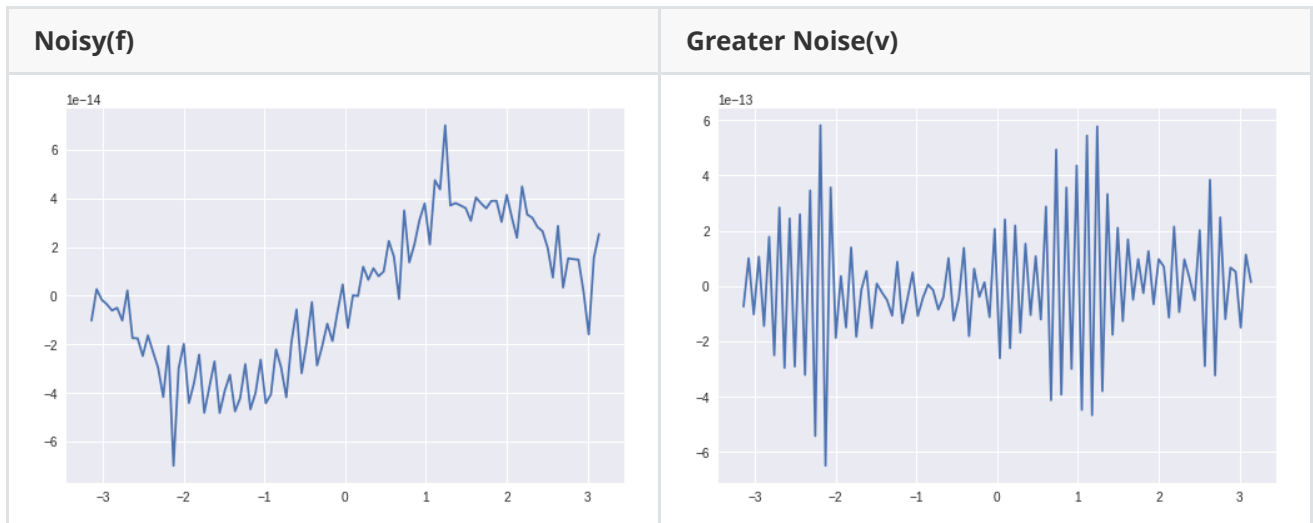
3. Smoothing

When a gaussian kernel is convolved with a noisy signal we get a smoothed version of the noisy signal.

For example we choose \mathbf{v} to be a noisy kernel and when smoothed with K which is a gaussian kernel we get $K \mathbf{v} = \mathbf{f}$.



If we choose a noisy \mathbf{f} and try to find the corresponding \mathbf{v} we end up with a signal that has greater noise than \mathbf{f} .



Our observations from section 2 seem to suggest that we are trying increase the cosine distance between the noisy signal(\mathbf{f}) and the signal with even greater noise \mathbf{v} i.e between \mathbf{f} and \mathbf{f}' .