14. GAUSSIAN PROCESS REGRESSION OF SPECTRAL LINES

14.1. *Theory*

The models that we propose to explain the observed spectral lines are never able to fit all the details that we observe. In some cases, we are interested in explaining the spectral line using a very simplified model because we are only interested in the general properties. In other cases, our models fail short in taking into account all the complications of the radiative transfer and our fit is always off.

To alleviate this issue, we propose to fit spectral lines (Stokes I only or the full Stokes vector) using a Gaussian process (GP). A Gaussian process is a very general prior distribution for functions and can easily fit very complex functions. They are extensively used for the non-parametric modeling of observations. The idea that we pursue is to use a parametric model that takes into account all we want to include about the spectral line formation and absorb the remaining components (usually unknown or neglected) in a non-parametric Gaussian process.

Following the standard recipes, regression with a Gaussian process begins by writing the following very general generative model:

$$y(x) = f(x) + \epsilon(x), \tag{106}$$

where $\epsilon(x)$ is a zero mean Gaussian variable with variance σ_n^2. Typically in the Gaussian process literature, we consider that the function $f(x)$ is very general and we learn it from the data, generating a non-parametric fit to the observations. The generative model and the statistical properties of $\epsilon(x)$ define the following likelihood for the set of observations $\mathbf{y} = (y_1, y_2, \ldots, y_N)$:

$$p(\mathbf{y}|\mathbf{f}, X) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_n^2 \mathbf{I}), \tag{107}$$

where X represent the x positions for all the measurements, while $\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \Sigma)$ represents a multi-variate Gaussian distribution for variable $\mathbf{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix Σ. As a prior for the function f, it is very advantageous to use the following Gaussian process:

$$p(\mathbf{f}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}, \mathbf{K}) \tag{108}$$

where $\boldsymbol{\mu}$ is the mean and $\mathbf{K}$ is the covariance matrix of the Gaussian process. There are several options for such covariance matrix and we later show some of them. It becomes advantageous to make the model hierarchical by making the mean and the covariance matrix depend on a set of hyperparameters. Concerning the mean, we make it depend on the set of parameters $\boldsymbol{\theta}$. These hyperparameters will be the thermodynamical and magnetic parameters of our model for the spectral line. Likewise, $\boldsymbol{\psi}$ refer to the parameters on which the covariance matrix depends. Using appropriate priors for these hyperparameters, the application of the Bayes theorem gives:

$$p(\mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}, X) = \frac{p(\mathbf{y}|\mathbf{f}, X) p(\mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\psi})}{p(\mathbf{y}|X)}, \tag{109}$$

where

$$p(\mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{K}(\boldsymbol{\psi})) p(\boldsymbol{\theta}, \boldsymbol{\psi}). \tag{110}$$

In principle, the full problem seems intractable since we should compute the posterior for $\mathbf{f}$, $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ and compute the appropriate marginalizations to get reliable estimations of the hyperparameters. The standard literature on Gaussian processes instead take a Type-II maximum-likelihood approach, in which $\mathbf{f}$ is marginalized:

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}, X) = \int d\mathbf{f} p(\mathbf{y}|\mathbf{f}, X) p(\mathbf{f}, \boldsymbol{\theta}, \boldsymbol{\psi}) \tag{111}$$

and then maximized with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. One has to be aware that this partially-Bayesian approach might be still prone to overfitting. The advantage of following this way is that the integral in Eq. (111) has an analytical expression under the assumption of a Gaussian likelihood and priors:

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}, X) = p(\boldsymbol{\theta}, \boldsymbol{\psi}) \int d\mathbf{f} \mathcal{N}(\mathbf{y}|\mathbf{f}, \sigma_n^2 \mathbf{I}) \mathcal{N}(\mathbf{f}|\boldsymbol{\mu}(\boldsymbol{\theta}), \mathbf{K}(\boldsymbol{\psi})). \tag{112}$$

If flat priors are assumed for $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$, we end up with:

$$p(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}, X) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}(\boldsymbol{\theta}), \sigma_n^2 \mathbf{I} + \mathbf{K}(\boldsymbol{\psi})), \tag{113}$$

a Gaussian with mean $\boldsymbol{\mu}(\boldsymbol{\theta})$ and covariance matrix $\mathbf{C} = \sigma_n^2 \mathbf{I} + \mathbf{K}(\boldsymbol{\psi})$. Maximizing the previos marginal distribution with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$ is equivalent to minimizing its negative logarithm, which is given by:

$$L = -\log p(\boldsymbol{\theta}, \boldsymbol{\psi}|\mathbf{y}, X) = \frac{1}{2} \log \det \mathbf{C}(\boldsymbol{\psi}) + \frac{1}{2} \left[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\psi})\right]^T \mathbf{C}^{-1}(\boldsymbol{\psi}) \left[\mathbf{y} - \boldsymbol{\mu}(\boldsymbol{\psi})\right] + \frac{N}{2} \log 2\pi. \tag{114}$$

The merit function to be minimized is composed of two terms. The quadratic term measures the quality of the fit, while the log-determinant term measures the complexity of the model. Therefore, the GP regression adapts to the quality of the data but trying not to overfit the noise. The optimization of Eq. (114) is roughly similar to what is done in standard inversion codes except for three reasons. First, the covariance matrix $\mathbf{C}$ is non-diagonal, contrary to what is assumed in standard inversion codes. It is non-diagonal not because we assume that the noise in different wavelength

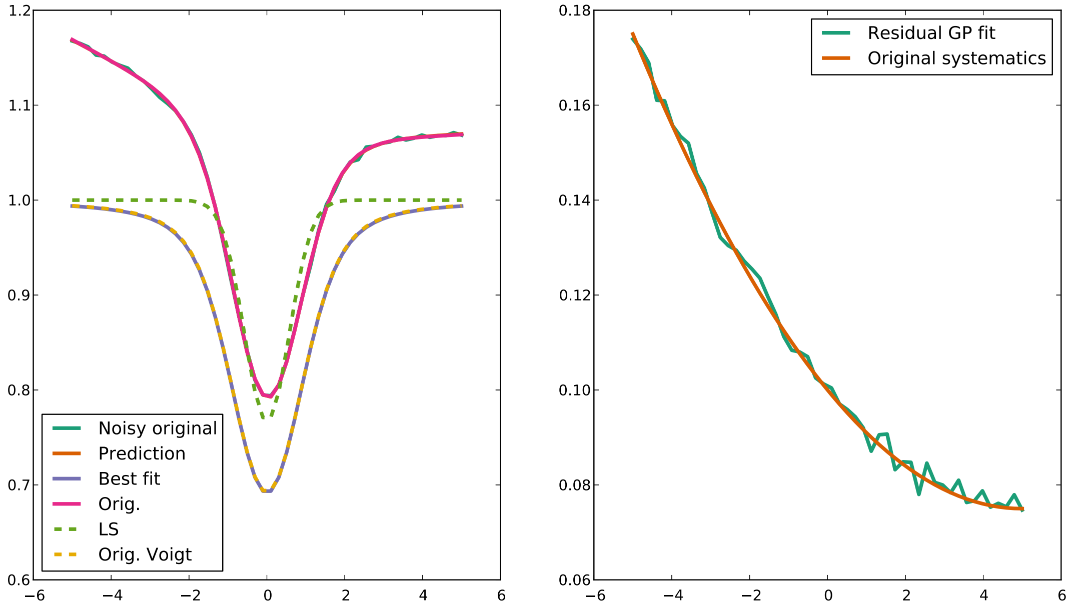

Figure 7. The left panel shows an example of fitting a Voigt function to noisy data (green curve) that is made of a Voigt function plus some systematics. The original Voigt function is shown in dashed yellow and the best fit obtained with the GP method is shown in solid violet curve. The fit obtained using a standard least-squares method is showin in dashed green. The right panel displays the original systematics (orange) and the inferred systematics.

points are correlated but because they are correlated under the non-parametric model used to explain everything that departs from our proposed model for the spectral line. Second, since the covariance matrix also depends on a set of hyperparameters, one has to augment the number of parameters of the model to include the parameters of the covariance matrix, making it more complex. Finally, there is an additionl term that depends on the log-determinant of the covariance matrix $\mathbf{C}$.

One can follow two approaches of increasing complexity. In the first one, if we fix the value of the hyperparameters $\boldsymbol{\psi}$, the problem is very similar to what we currently do for inversion codes, but with a non-diagonal covariance matrix. Options for the covariance matrix (combinations of the following functions are also possible as covariance matrices) are (see Rasmussen & Williams 2006):

- Constant: $K_{ij} = C$
- Linear: $K_{ij} = x_i x_j$
- Gaussian noise: $K_{ij} = \sigma^2 \delta_{ij}$
- Squared exponential:
$$K_{ij} = \sigma^2 \exp\left(-\frac{(x_i - x_j)^2}{2l^2}\right). \tag{115}$$
This is one of the most used covariance matrices and is specially suitable for smooth functions.
- Matérn:
$$K_{ij} = \frac{2^{1-v}}{\Gamma(v)} \left(\frac{\sqrt{2v}|x_i - x_j|}{l}\right)^v K_v \left(\frac{\sqrt{2v}|x_i - x_j|}{l}\right) \tag{116}$$
This is also a very general covariance matrix that is suitable for slightly less smooth functions.
- Periodic:
$$K_{ij} = \exp\left(-\frac{2\sin^2 \frac{d}{2}}{l^2}\right). \tag{117}$$
This covariance matrix is adequate for periodic signals.

In the second approach, the full set of hyperparameters should be obtained. For this reason, it is interesting to note that Eq. (114) can be minimized noting that the derivatives with respect to the hyperparameters are given by:

$$\begin{aligned} -\frac{\partial L}{\partial \psi_k} &= \frac{1}{2}\mathrm{tr}\left(\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \psi_k}\right) + \frac{1}{2}\left[\mathbf{y}-\boldsymbol{\mu}\right]^T \frac{\partial \mathbf{C}}{\partial \psi_k}\mathbf{C}^{-1}\frac{\partial \mathbf{C}}{\partial \psi_k}\left[\mathbf{y}-\boldsymbol{\mu}\right] \\ -\frac{\partial L}{\partial \theta_k} &= \left[\mathbf{y}-\boldsymbol{\mu}\right]^T \mathbf{C}^{-1}\frac{\partial \boldsymbol{\mu}}{\partial \theta_k}, \end{aligned} \tag{118}$$

which are fast to compute, given that $\mathbf{C}^{-1}$ is already needed for the computation of the merit function. Once a suitable covariance matrix is chosen, the derivatives $\partial \mathbf{C}/\psi_k$ are easy to compute. The remaining derivatives $\partial \boldsymbol{\mu}/\partial \theta_k$ are also needed and, once a suitable model to explain the spectral line is proposed, it should be straightforward. In simple models based on Milne-Eddington atmospheres these derivatives can be computed analytically. In more complex models with gradients in the physical conditions, one has to rely on the computation of the response functions, like in current inversion codes. It is important to note that the optimization of Eq. (114) is not convex, so that we cannot claim that a local minimum is the global minimum. Therefore, care must be taken during the optimization, like in any standard inversion code.

14.2. *Example*

One of the interesting applications of the previous formalism is to take into account unknown systematics while fitting spectral lines. We consider fitting a spectral line that is given by a Voigt profile and that is perturbed by an additive systematics:

$$I(\lambda) = 1 - dH\left(\frac{\lambda-\lambda_0}{\delta}, a\right) + f(\lambda), \tag{119}$$

where $d = 0.5$ is the depth of the line, λ_0 is the line center, $\delta = 1$ is the width of the line, $a = 0.5$ is the damping parameter and $H(v,a)$ is the Voigt profile that is assumed to be normalized to unit area. In our example, we used $f(\lambda) = 0.001(\lambda-\lambda_0)^2 - 0.01(\lambda-\lambda_0) + 0.1$. We used this model as the mean of the GP, that depends on the hyperparameters $\boldsymbol{\theta} = (d, \delta, a)$ and propose a squared-exponential covariance matrix that depends on the parameters $\boldsymbol{\psi} = (\sigma, l)$. We optimize the merit function of Eq. (114) with respect to the five hyperparameters and find the results displayed in the left panel of Fig. 7. The green solid line displays the observed profile, that is perturbed by the systematics. The green dashed line is the Voigt profile fitted with a standard least-squares code. The yellow dashed line is the original Voigt profile without systematics. The violet curve represents the fit with the GP approach and the pink curve is the prediction of the GP including the systematics, that perfectly fits the observations. The right panel shows the residuals between the GP fit and the observations, together with the original systematics. We see that the covariance of the GP is able to correctly capture these systematics. Concerning the parameters of the fits, we find:

$$\begin{aligned} d_{\mathrm{GP}} &= 0.4899 - w_{\mathrm{GP}} = 1.0052 - a_{\mathrm{GP}} = 0.4870 \\ d_{\mathrm{LS}} &= 0.2344 - w_{\mathrm{LS}} = 0.8171 - a_{\mathrm{LS}} = 1.1433e - 07, \end{aligned} \tag{120}$$

demonstrating that the parameters of the fitted GP closely resemble the original ones.