TIME-DOMAIN ASTRONOMY

Lectures 9: Gaussian Processes

Stefano Covino
INAF / Brera Astronomical Observatory

Gaussian Processes

- A GP is a type of *stochastic process* based on the Gaussian probability distribution.
- The formal definition of a GP is that the joint probability distribution over any finite sample $y = \{y_i\}_{i=1,\ldots,N}$ from the GP is a multi-variate Gaussian:

$$p(y) = N(\mathbf{m}, \mathbf{K}),$$

  - where $\mathbf{m}$ is the mean vector and $\mathbf{K}$ the covariance matrix.

Gaussian Processes

- The elements of the mean vector and covariance matrix are given by the *mean function m* and the *covariance function k*, respectively:

$$m_i = m(x_i, \boldsymbol{\theta}),$$
$$K_{ij} = k(x_i, x_j, \varphi).$$

- where x_i is the set of inputs (independent variables) corresponding to the i^{th} sample.

- For time-series data, the inputs usually include, but aren't necessarily restricted to, the time t_i.
- The covariance function is also known as the kernel function, and it is a fundamental ingredient of a GP model,

Gaussian Processes

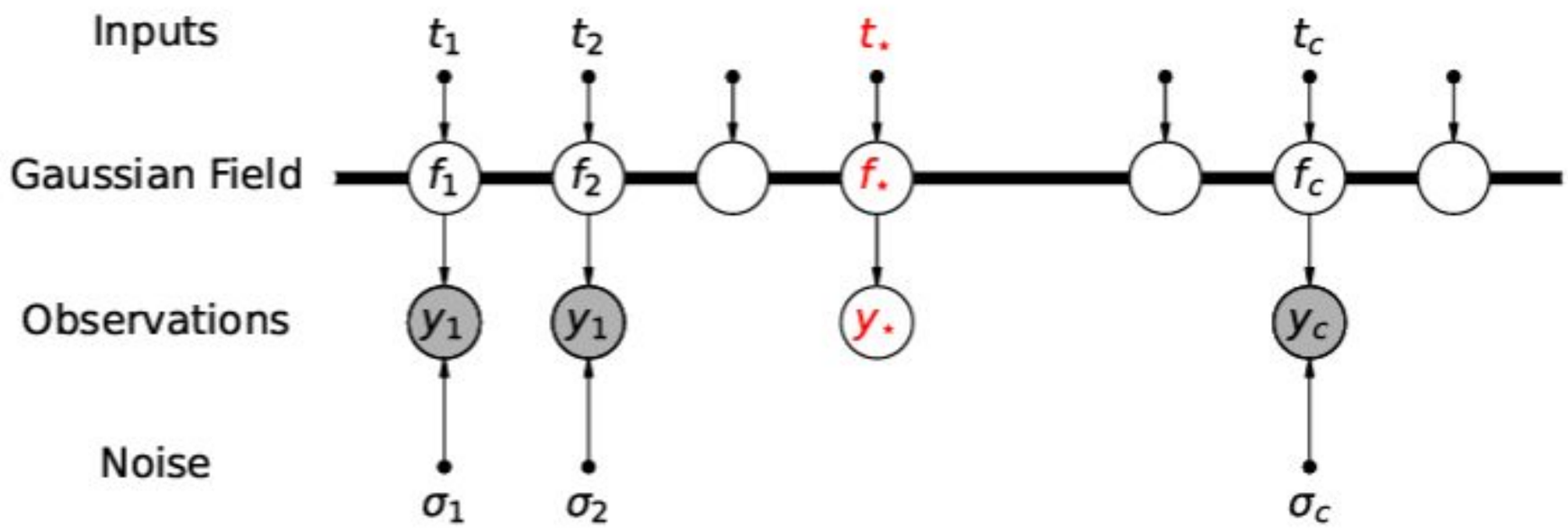


Figure 4

Probabilistic Graphical Model (chain graph) for GP regression (adapted from Figure 2.3 of Rasmussen & Williams 2006) given observations $\mathbf{y}$ at inputs $\mathbf{t}$, and a prediction at test input $t_\star$. Dots represent fixed variables (inputs), grey circles represent observed variables white circles represent unknown (latent) variables. The thick thick horizontal bar indicates a set of fully connected nodes (meaning that every f depends on every t). On the other hand, each observed output y_i depends only on the corresponding f_i and σ_i, and is conditionally independent of the other variables). In this example, the measurement uncertainties $\boldsymbol{\sigma}$ and the GP hyper-parameters (not shown) are treated as known (fixed).

Gaussian Processes

- The parameters θ and φ of the mean and covariance function are known as *hyper-parameters* of the GP.

- Strictly speaking, the parameters of the GP are the (infinitely many) unknown functions that share the specified mean vector and covariance matrix and could have given rise to the observations. However, these parameters are always marginalised over: we never explicitly deal with the individual functions.

- GPs are therefore a type of Hierarchical Bayesian Model (HBM).

Other Non-Gaussian Processes?

- In principle, it is possible to construct and use stochastic process models based on other distributions, GPs are by far the most popular, for two main reasons:

  - The first is Central Limit theorem: it implies that the assumption of Gaussianity is often at least approximately correct.

  - The second is that Gaussian distributions obey simple mathematical identities for marginalisation and conditioning, that enable inference with GPs.

A pragmatic approach to GPs

- GP regression (GPR) can be thought of as a generalisation of least-squares regression, allowing for correlated noise (or signals) in the data.
- Conversely, least-squares regression, as traditionally presented, is a special case of GPR, where the covariance matrix is assumed to be purely diagonal, and the variances associated with each observation are known *a priori*.

Least-square Regression

- Let's consider N observations of a variable $\mathbf{y}=\{y_i\}_{i=1,\ldots,N}$, taken at times $\mathbf{t}=\{t_i\}$, with associated measurement uncertainties $\boldsymbol{\sigma}=\{\sigma_i\}$.
- We wish to compare these to a model function $m(t,\theta)$ controlled by parameters $\boldsymbol{\theta}=\{\theta_j\}_{j=1,\ldots,M}$.
- In least-squares regression, we minimize the quantity:

$$\chi^2 \equiv \sum_{i=1}^{N} (y_i - m_i)^2 / \sigma_i^2,$$

  - where $m_i \equiv m(t_i, \boldsymbol{\theta})$, with respect to $\boldsymbol{\theta}$.

Least-square Regression

- The χ^2 minimization come from the following considerations:

- Let us assume that the observations are given by $y_i = m(t_i, \boldsymbol{\theta}) + \varepsilon_i$.
  - where ε_i is the measurement error, or noise, on the i^{th} observation.

- Let us also assume that ε_i is drawn from a Gaussian distribution with mean 0 and variance σ^2_i:

$$p(\epsilon_i) = \mathcal{N}(0, \sigma_i^2) \equiv \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left(-\frac{\epsilon_i^2}{2\sigma_i^2}\right),$$

Least-square Regression

- Then the *likelihood* for the i^{th} observation is simply:

$$\mathcal{L}_i(\boldsymbol{\theta}) \equiv p(y_i|\boldsymbol{\theta}) = \mathcal{N}(m_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(y_i - m_i)^2}{2\sigma_i^2}\right].$$

- We also assume that the noise is uncorrelated, or white, meaning that the ε_i's are drawn independently from each other from their respective distributions.

- Then, the likelihood for the whole dataset **y** is merely the product of the likelihoods for the individual observations ($m=\{m_i\}_{i=1,\ldots,N}$):

$$\mathcal{L}(\boldsymbol{\theta}) \equiv p(\mathbf{y}|\boldsymbol{\theta}) = \prod_{i=1}^{N} \mathcal{L}_i = \prod_{i=1}^{N} \left\{ \frac{1}{\sqrt{2\pi}\sigma_i} \exp\left[-\frac{(y_i - m_i)^2}{2\sigma_i^2}\right] \right\},$$

Least-square Regression

- One can readily see that:

$$\ln L = \text{constant} - 0.5\chi^2,$$

  - where the constant depends only on the σ's.

- Thus, if the σ's are known, maximizing L is equivalent to minimizing χ^2.

- In other words, least-squares regression yields the Maximum Likelihood Estimate (MLE) of the parameters under the assumption of white, Gaussian noise with known variance.

Least-square Regression → GPR

- Let us now re-write the likelihood in matrix form:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\phi}) = \frac{1}{\sqrt{2\,\pi\,|\mathbf{K}|}} \exp\left(-(\mathbf{y}-\mathbf{m})^{\mathrm{T}}\mathbf{K}^{-1}(\mathbf{y}-\mathbf{m})\right),$$

- where the mean vector $\mathbf{m}$ has elements $m_i = m(t_i, \boldsymbol{\theta})$ and $\mathbf{K}$ is a purely diagonal (N,N) matrix with elements $K_{ij} = \delta_{ij}\sigma^2_i$ (δ_{ij} being the discrete Kronecker delta function).

- $\mathbf{K}$ is the *covariance matrix* of the model!

Least-square Regression → GPR

- Let us now allow a more flexible covariance model:

$$K_{ij} = k(t_i, t_j, \boldsymbol{\phi}) + \delta_{ij}\sigma_i^2,$$

- k is the *covariance function*, or *kernel function*, controlled by parameters φ. The result is a GP!

- Depending on the choice of kernel function and parameters, the covariance matrix can now have non-zero off-diagonal elements, allowing us to explicitly model correlated noise or stochastic signals in the data.
- The kernel function encodes our beliefs about the stochastic, or random, element of the model, in just the same way as the mean function encodes our beliefs about the deterministic component of the model.

Standard Kernel Functions

- The kernel function can be any positive scalar function that gives rise to a positive semi-definite covariance matrix over the input domain.
- Some popular kernel functions are listed below:

Name	Representation[a]
Constant	α^2
Squared Exponential[b]	$e^{-(\tau/\lambda)^2/2}$
Exponential[c]	$e^{-\tau/\lambda}$
Matérn-3/2	$\left(1+\sqrt{3}\tau/\lambda\right)e^{-\sqrt{3}\tau/\lambda}$
Matérn-5/2	$\left(1+\sqrt{5}\tau/\lambda+5(\tau/\lambda)^2/3\right)e^{-\sqrt{5}\tau/\lambda}$
Rational quadratic	$\left(1+\frac{\tau^2}{2\gamma\lambda^2}\right)^{-\gamma}$
Cosine	$\cos 2\pi\tau/\lambda$
Sine Squared Exponential	$\exp\left(-\Gamma\sin^2\pi\tau/\lambda\right)$
Stochastic Harmonic Oscillator[d]	$\cos\left(\sqrt{1-\beta^2}\frac{\tau}{\lambda}\right)+\frac{\beta}{\sqrt{1-\beta^2}}\sin\left(\sqrt{1-\beta^2}\frac{\tau}{\lambda}\right)$

[a] in each case, τ is defined as $\tau = |t_i - t_j|$, and Greek letters indicate hyper-parameters; [b] "radial basis function"; [c] "Ornstein–Uhlenbeck", "damped random walk" or "Matérn-1/2"; [d] Foreman-Mackey et al. (2017).

Kernel Functions

- More kernel functions are often constructed using products and sums of the standard kernels.
- Besides, addition and multiplication, other operations can be used to impose structure on the standard kernels.
  - For example, linear operations like scalar multiplication, more general affine transformations, differentiation, or integration can all be used to develop new kernel functions.
- A general rule is that the error associated to observation j is typically larger (or even much larger) than its variance due to the covariance matrix (e.g. red-noise) contribution.

Making predictions by GPs

- Given some existing observations $\mathbf{y}$, taken at times $\mathbf{t}$, we can make predictions at some new set of times $\mathbf{t}_*$, i.e., computing $p(\mathbf{y}_*|\mathbf{y})$, the conditional probability distribution for $\mathbf{y}_*$ given $\mathbf{y}$.

- This is often called *predictive distribution*, because it is often used to extrapolate a time-series dataset forwards.

- The predictive distribution is also Gaussian, and its mean and covariance are given by simple analytic relations:

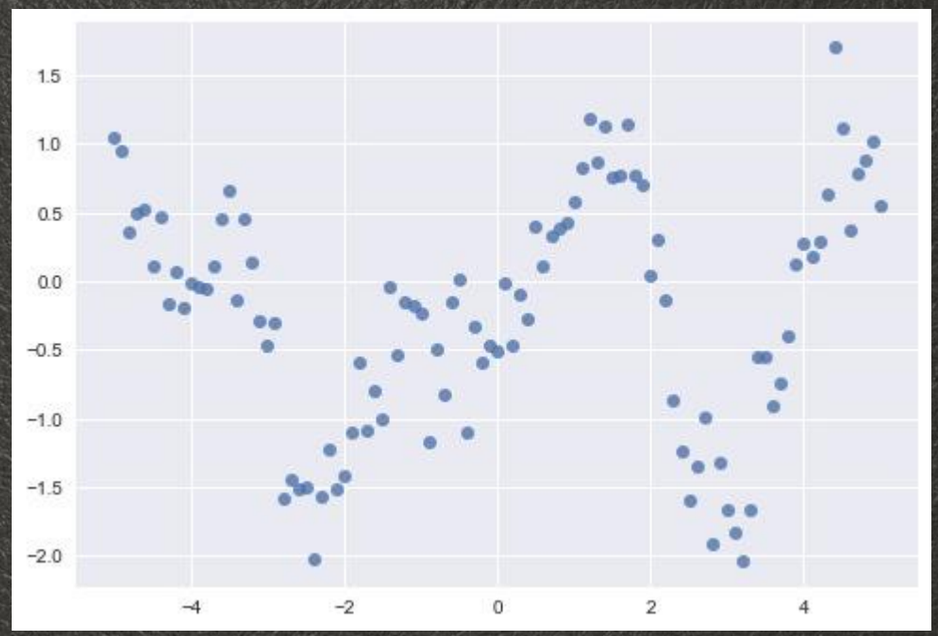
$$\mathbf{f}_\star = \mathbf{m}_\star + \mathbf{K}_\star^{\mathrm{T}}\,\mathbf{K}^{-1}\,(\mathbf{y}-\mathbf{m}) \quad \text{and} \quad \mathbf{C}_\star = \mathbf{K}_{\star\star} - \mathbf{K}_\star^{\mathrm{T}}\,\mathbf{K}^{-1}\,\mathbf{K}_\star,$$

  - where $\mathbf{m}_* \equiv m(\mathbf{t}_*,\boldsymbol{\theta})$, $\mathbf{K}_* \equiv k(\mathbf{t},\mathbf{t}_*,\varphi)$ and $\mathbf{K}_{**} \equiv k(\mathbf{t}_*,\mathbf{t}_*,\varphi)$.

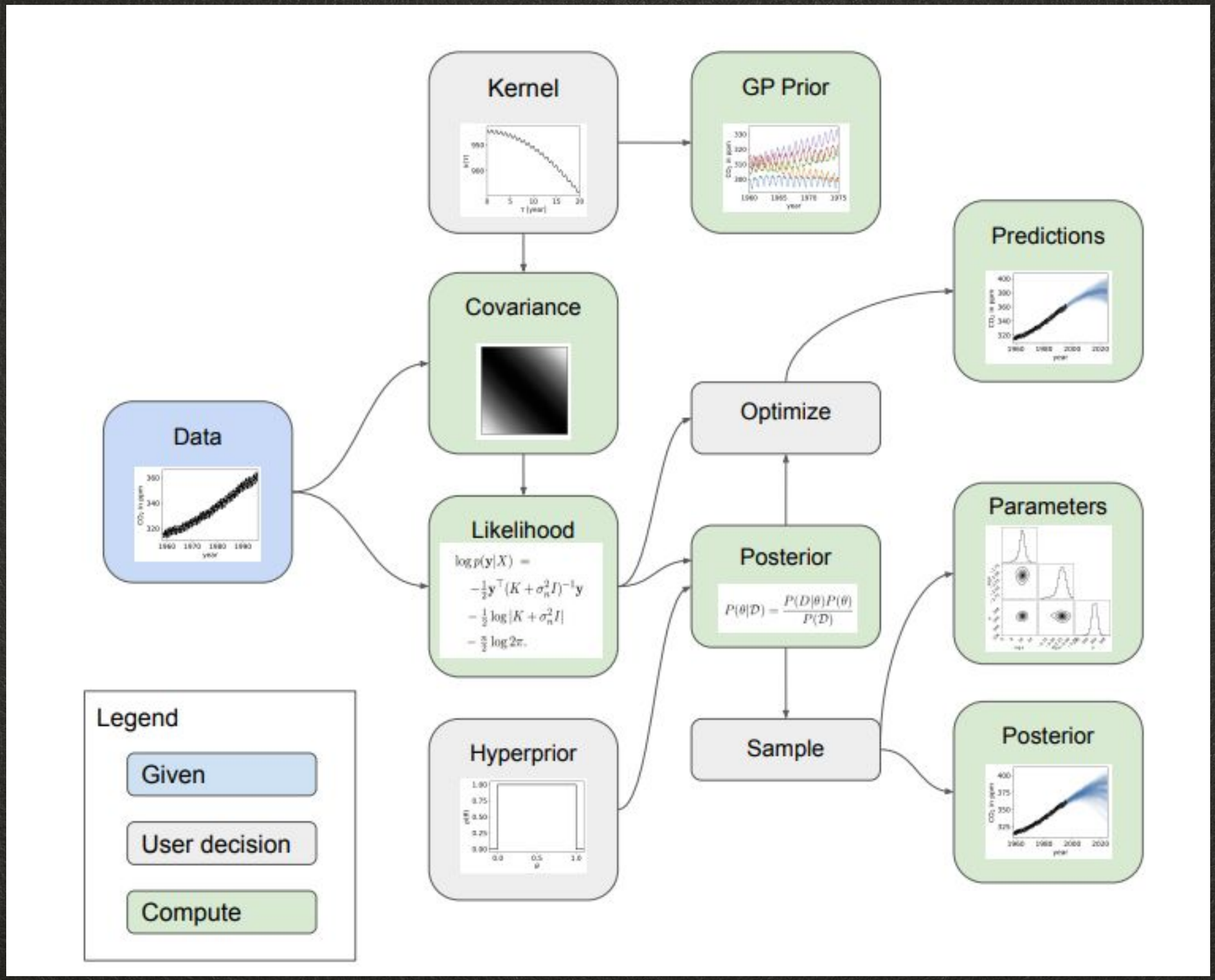
Exercise

Useful notebook:

1. SamplingwithaGP



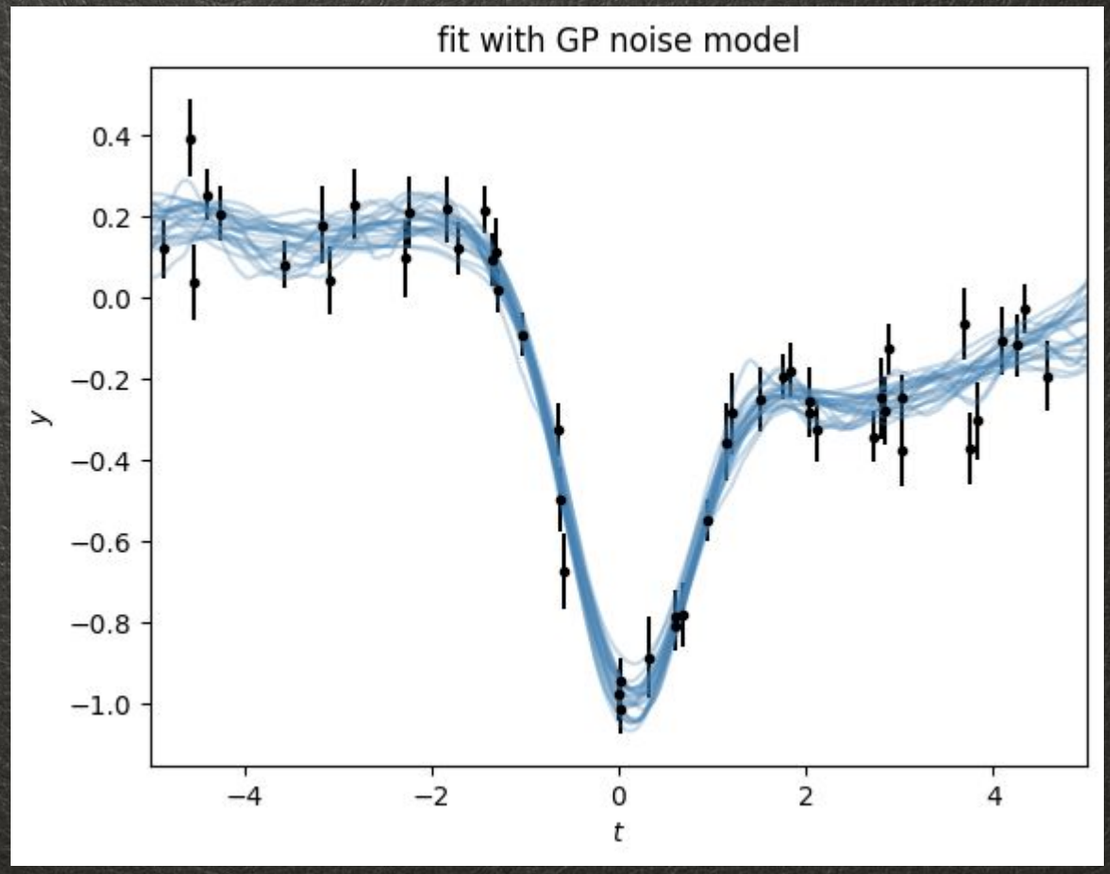
GPR Inference workflow



Exercise

Useful notebook:

1. CorrelatedNoise
2. MultipleDelayedTimeSeries



Multivariate GPs

- Since we focus on time domain astronomy, we primarily consider univariate GPs with just time as the input coordinate.

- As a matter of fact, most of our discussion applies to multivariate datasets without any change.

- However, when working with univariate GPs, it is relatively straightforward and unambiguous to compute the "distance" between two points t_i, and t_j: $\rho(t_i,t_j)=\tau=|t_i-t_j|$.

- For multivariate inputs $\mathbf{x}$, more care must be taken to define a sensible distance metric $\rho(\mathbf{x}_i,\mathbf{x}_j)$.

Multivariate GPs

- A common class of multivariate GP model for time domain astronomy are datasets with multiple parallel covariant time series.
  - For example, multi-band time series produced by surveys like PanSTARRS or LSST, or radial velocity time series with parallel activity indicators.

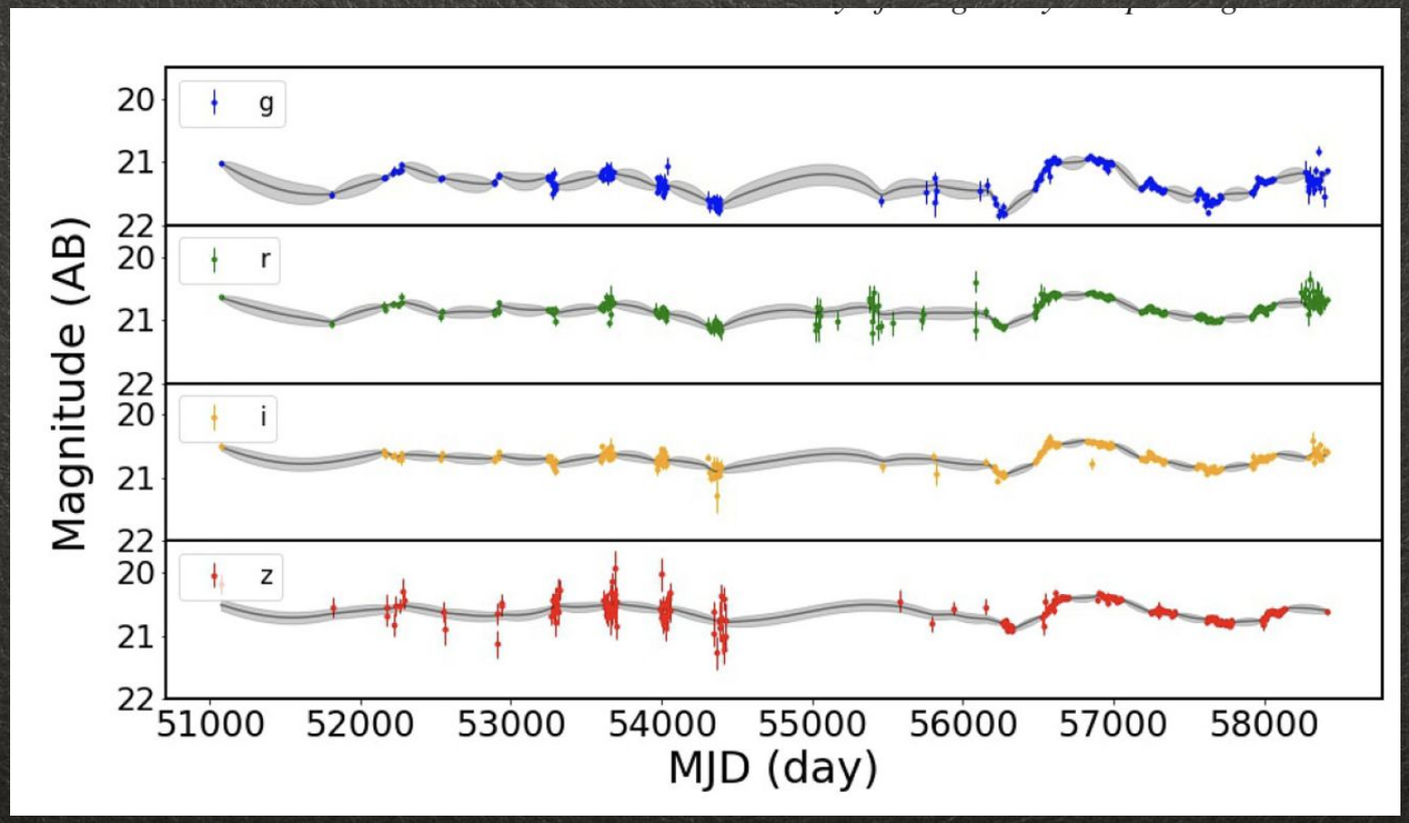
- One useful way to specify these datasets are to define the inputs as $\mathbf{x_i} \equiv (t_i, l_i)$, where t_i the time of the i-th observation and l_i is a"label" for which time series the i-th observation is drawn from.

Multivariate GPs

- In this case, one common choice of kernel function is:

$$k(\boldsymbol{x}_i, \boldsymbol{x}_j; \boldsymbol{\phi}) = \boldsymbol{a}_{\ell_i}^{\mathrm{T}} \boldsymbol{a}_{\ell_j} k_0(t_i, t_j; \boldsymbol{\phi}) \quad ,$$

  - where $k_0(t_i,t_j;\varphi)$ is a standard one-dimensional kernel, and the set of $\{\mathbf{a_d}\}^{D}_{d=1}$ are also hyper-parameters of the model.



Covino et al. (2022)

REFERENCES AND DEEPENING

Susanne Aigrain

Daniel Foreman-Mackey

Gaussian Process regression for astronomical time-series

Suzanne Aigrain,[1] and Daniel Foreman-Mackey[2]

[1]Department of Physics, University of Oxford, Oxford, UK, OX1 3RH; email: suzanne.aigrain@physics.ox.ac.uk

[2]Center for Computational Astrophysics, Flatiron Institute, New York, USA, NY 10010; email: dforeman-mackey@flatironinstitute.org

Annu. Rev. Astron. Astrophys. 2022. AA:1–40

https://doi.org/10.1146/TBD

Keywords

Gaussian process regression, astronomy data analysis, time-series analysis, time domain astronomy, astrostatistics techniques, computational methods

Abstract

The last two decades have seen a major expansion in the availability, size, and precision of time-domain datasets in astronomy. Owing to their unique combination of flexibility, mathematical simplicity and comparative robustness, Gaussian Processes (GPs) have emerged recently as the solution of choice to model stochastic signals in such datasets. In this review we provide a brief introduction to the emergence of GPs in astronomy, present the underlying mathematical theory, and give practical advice considering the key modelling choices involved in GP regression. We then review applications of GPs to time-domain datasets in the astrophysical literature so far, from exoplanets to active galactic nuclei, showcasing the power and flexibility of the method. We provide worked examples using simulated data, with links to the source code, discuss the problem of computational cost and scalability, and give a snapshot of the current ecosystem of open source GP software packages. Driven by further algorithmic and conceptual advances, we expect that GPs will continue to be an important tool for robust and interpretable time domain astronomy for many years to come.