GAUSSIAN PROCESS REGRESSION

Susheela Singh

North Carolina State University

November 30, 2016

OVERVIEW

Motivation

Prediction with Gaussian processes is not a new idea. It has roots that date back to Kolmogorov in the 1940s and applications to multivariate regression as early as the 1960s.

Motivation

Prediction with Gaussian processes is not a new idea. It has roots that date back to Kolmogorov in the 1940s and applications to multivariate regression as early as the 1960s.

- ARMA models in time series analysis
- "Kriging" in geostatistical models
- Regression splines

MOTIVATION

Prediction with Gaussian processes is not a new idea. It has roots that date back to Kolmogorov in the 1940s and applications to multivariate regression as early as the 1960s.

- ARMA models in time series analysis
- "Kriging" in geostatistical models
- Regression splines

Gaussian process regression is a "less" parametric tool for supervised learning.

WHAT IS A GAUSSIAN PROCESS?

A stochastic process, $Y(\mathbf{x})$, is a Gaussian process (GP) if it generates data such that any finite subset of the range of the process follows a multivariate Gaussian distribution.

Specification

Because the joint distribution of $Y(\mathbf{x}_1), Y(\mathbf{x}_2), \ldots, Y(\mathbf{x}_n)$ is multivariate Gaussian, we need only specify the mean and the covariance functions:

- $\mathbb{E}[Y(\mathbf{x})] = m(\mathbf{x})$
- $\mathbb{E}[\{Y(\mathbf{x}) - m(\mathbf{x})\}\{Y(\mathbf{x}') - m(\mathbf{x}')\}^T] = k(\mathbf{x}, \mathbf{x}')$.

Then, we write $Y(\mathbf{x}) \sim \mathcal{GP}\left(m(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')\right)$.

Regression Model

Suppose that we observe $y_1, \ldots, y_n$, which are measured without error (for now).

We believe there is an underlying process $f(\mathbf{x})$ such that

$$y = f(\mathbf{x}).$$

Our goal is to estimate $f(\mathbf{x})$. To do so, we will assume that $f(\mathbf{x}) \sim \mathcal{GP}\left(0,\ k(\mathbf{x}, \mathbf{x}')\right)$.

SPECIFYING A COVARIANCE FUNCTION

The covariance function, $k(\mathbf{x}, \mathbf{x}')$, can be any function that generates a non-negative definite covariance matrix for any finite set of points $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

Specifying a Covariance Function

The covariance function, $k(\mathbf{x}, \mathbf{x}')$, can be any function that generates a non-negative definite covariance matrix for any finite set of points $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

- Seems general, but these functions are tricky to find!
- Usually rely on families that are already well-studied.

COMMON COVARIANCE FUNCTIONS

- Constant:

$$k(\mathbf{x}, \mathbf{x}') = v_0$$

- Gaussian Noise:

$$k(\mathbf{x}, \mathbf{x}') = v_0 \; \delta_{\mathbf{x}, \mathbf{x}'}$$

- Squared Exponential:

$$k(\mathbf{x}, \mathbf{x}') = v_0 \exp \left[-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{2\ell^2} \right]$$

- Many options: Ornstein-Uhlenbeck, Matérn, periodic, stationary and isotropic covariance functions from spatial statistics, etc.

A More General Covariance Function

Williams and Rasmussen (1996) propose a very general covariance function that is flexible and works well in practice.

For $\mathbf{x} = (x_1, \ldots, x_p)$ and $\mathbf{x}' = (x'_1, \ldots, x'_p)$,

$$k(\mathbf{x}, \mathbf{x}') = v_0 \exp\left[-\frac{1}{2}\sum_{\ell=1}^{p} \alpha_\ell (x_\ell - x'_\ell)^2\right] + \beta_0 + \beta_1 \mathbf{x}^T \mathbf{x}' + v_1 \; \delta_{\mathbf{x},\mathbf{x}'}.$$

A More General Covariance Function, Explained

$$k(\mathbf{x}, \mathbf{x}') = v_0 \exp\left[-\frac{1}{2}\sum_{\ell=1}^{p} \alpha_\ell (x_\ell - x'_\ell)^2\right] + \beta_0 + \beta_1 \mathbf{x}^T \mathbf{x}' + v_1\, \delta_{\mathbf{x},\mathbf{x}'}.$$

- Nearby input values will have highly correlated outputs.
- Very similar to squared exponential, but allows a different level of smoothing for each input dimension.
- v_0 controls the overall scale of local correlations.

A More General Covariance Function, Explained

$$k(\mathbf{x}, \mathbf{x}') = v_0 \exp\left[-\frac{1}{2}\sum_{\ell=1}^{p} \alpha_\ell (x_\ell - x'_\ell)^2\right] + \beta_0 + \beta_1 \mathbf{x}^T \mathbf{x}' + v_1\, \delta_{\mathbf{x},\mathbf{x}'}.$$

- β_0 allows for bias, i.e. correlation not explained by inputs.
- β_1 allows for a linear contribution to the covariance.

A More General Covariance Function, Explained

$$k(\mathbf{x}, \mathbf{x}') = v_0 \exp\left[-\frac{1}{2}\sum_{\ell=1}^{p} \alpha_\ell (x_\ell - x'_\ell)^2\right] + \beta_0 + \beta_1 \mathbf{x}^T \mathbf{x}' + v_1\ \delta_{\mathbf{x},\mathbf{x}'}.$$

- Accounts for noise or measurement error in the data.
- v_1 controls the variance of the noise.

NOW WHAT?

We have specified our Gaussian process, but how do we use that to actually *perform* regression?

PREDICTION USING GAUSSIAN PROCESSES

For simplicity, suppose we want to predict the value at a single new point, $y_* = f(\mathbf{x}_*)$.

As always, first some notation:

- Let $\boldsymbol{y} = (y_1, \ldots, y_n)$ be the observed values.
- Let $\boldsymbol{K}$ be the $n \times n$ covariance matrix where $[\boldsymbol{K}]_{ij} = k(\mathbf{x}_i, \mathbf{x}_j)$.
- Let $\boldsymbol{K}_*$ be the $1 \times n$ vector, $\boldsymbol{K}_* = [k(\mathbf{x}_*, \mathbf{x}_1) \ldots k(\mathbf{x}_*, \mathbf{x}_n)]$.
- Let $K_{**} = k(\mathbf{x}_*, \mathbf{x}_*)$ be the scalar variance at the new point.

Prediction using Gaussian Processes

By the definition of Gaussian process, we know that the observed values and the desired predicted value have a joint multivariate normal distribution,

$$\begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix} \sim \mathcal{N}_{n+1}\left(\begin{bmatrix} \mathbf{0}_n \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{K} & \boldsymbol{K}_*^T \\ \boldsymbol{K}_* & K_{**} \end{bmatrix}\right).$$

Prediction using Gaussian Processes

By the definition of Gaussian process, we know that the observed values and the desired predicted value have a joint multivariate normal distribution,

$$\begin{bmatrix} \boldsymbol{y} \\ y_* \end{bmatrix} \sim \mathcal{N}_{n+1}\left(\begin{bmatrix} \boldsymbol{0}_n \\ 0 \end{bmatrix}, \begin{bmatrix} \boldsymbol{K} & \boldsymbol{K}_*^T \\ \boldsymbol{K}_* & K_{**} \end{bmatrix} \right).$$

We know the distribution of $y_*|\boldsymbol{y}$ exactly, so our best guess for y_* is simply the mean of this conditional distribution

$$\hat{y}_* = \boldsymbol{K}_* \boldsymbol{K}^{-1} \boldsymbol{y}.$$

APPLICATION TO MODELING LIGHTCURVES

The authors' goal is to develop a new set of measures that can enhance classification of objects based on lightcurves.

To that end, they want to estimate the "true" lightcurve based on sparse-ish observations and base their measures on that estimated function.

Application to Modeling Lightcurves

The authors' goal is to develop a new set of measures that can enhance classification of objects based on lightcurves.

To that end, they want to estimate the "true" lightcurve based on sparse-ish observations and base their measures on that estimated function.

Because the authors estimated each lightcurve individually, we will go through this process for a single curve to demonstrate how Gaussian process regression works in this setting.

DATA

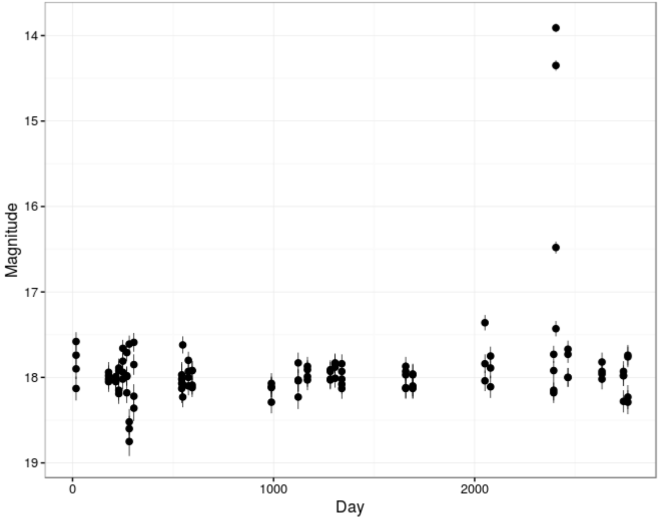

Figure: CSS111103:230309+400608, a Flare star

Regression Model

Using time, t, as the indexing variable for our proposed process and accounting for measurement error we posit that the process generating the observed magnitudes is of the form

$$y = f(t) + \mathcal{N}(0, \sigma_n^2),$$

where $f(t) \sim \mathcal{GP}\left(m(t), k(t, t')\right)$.

It remains to specify the functional forms of $m(t)$ and $k(t, t')$.

Specification

In this case, it seems that assuming $m(t) \equiv 0$ is inappropriate. Following the authors, we set $m(t) = 17.99$, the median observed magnitude in the data.

We use the squared exponential covariance kernel for $k(t, t')$. Recall that we have additional error in the observations caused by measurement error.

So, the covariance kernel for the *observed* magnitudes is

$$k_y(t, t') = \sigma_f^2 \exp\left[-\frac{1}{2\ell^2}(t - t')^2\right] + \sigma_n^2\, \delta_{t,t'}.$$

PARAMETER SPECIFICATION

As the authors suggest, we set the parameter values as

- $\sigma_f^2 = 0.27$, the median observed variance in the magnitudes of non-variable objects.
- $\sigma_n^2 = 0.01$, the mean value of measurement error in the data.
- $\ell = 140$ days.

FITTED CURVE

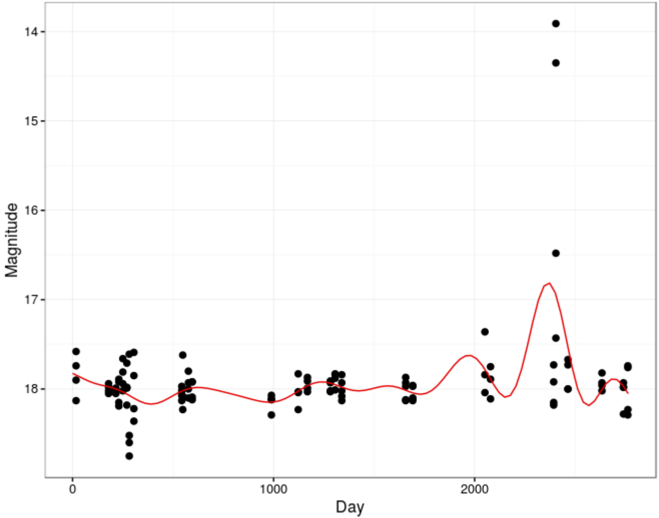

Figure: CSS111103:230309+400608, with smoothed curve in red

References

- Ebden, M. (2008), Gaussian processes for regression: a quick introduction.
- Faraway, J., Mahabal, A., Sun, J., Wang, X.-F., Wang, Y. G. and Zhang, L. (2016), Modeling lightcurves for improved classification of astronomical objects.
- Williams, C.K.I. and Rasmussen, C.E. (1996), Gaussian processes for machine learning.
- Williams, C.K.I. (1997), Prediction with gaussian processes: from linear regression to linear prediction and beyond.

Thank you!