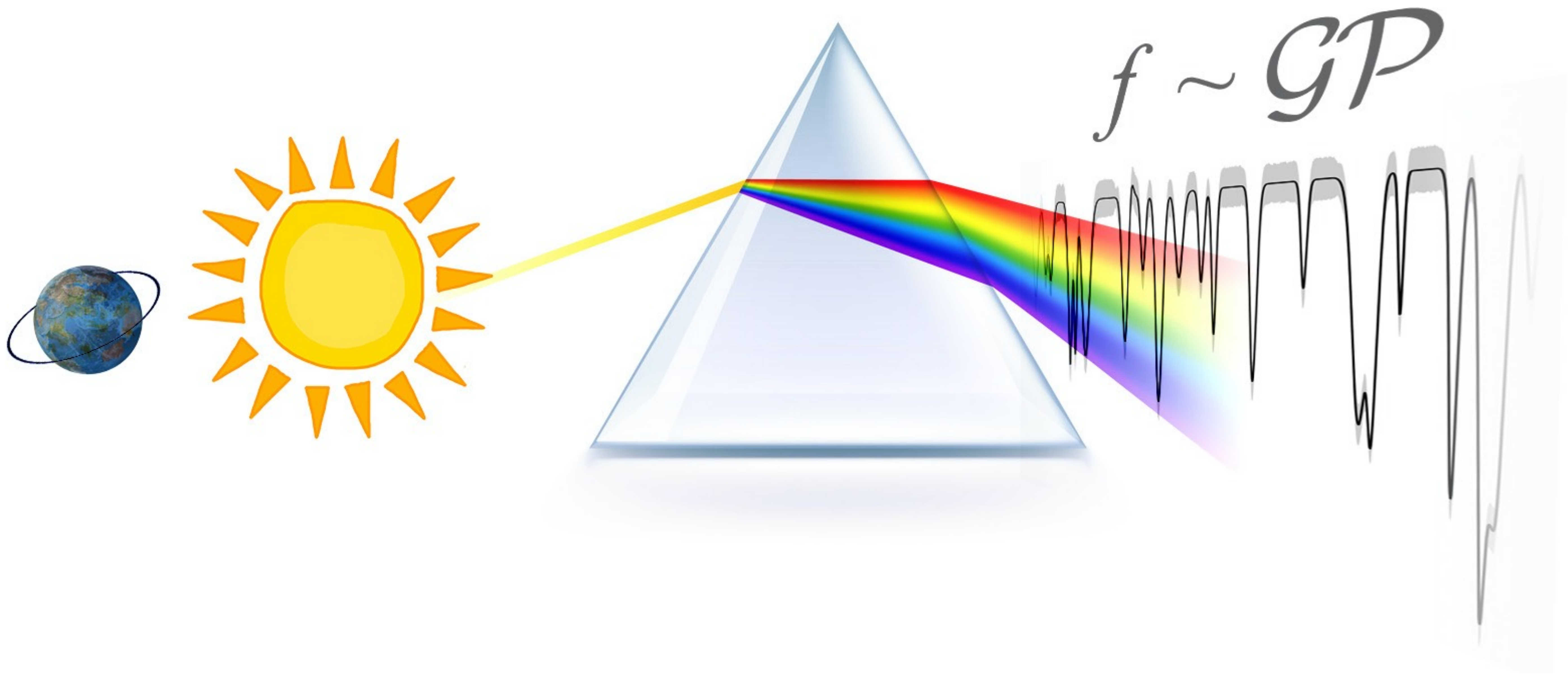
GAUSSIAN PROCESSES

THEIR POWER AND LIMITATIONS



Vinesh Maguire-Rajpaul
Sagan Summer Workshop 2020

Royal Astronomical Society
UNIVERSITY OF CAMBRIDGE
Cavendish Laboratory
Emmanuel College

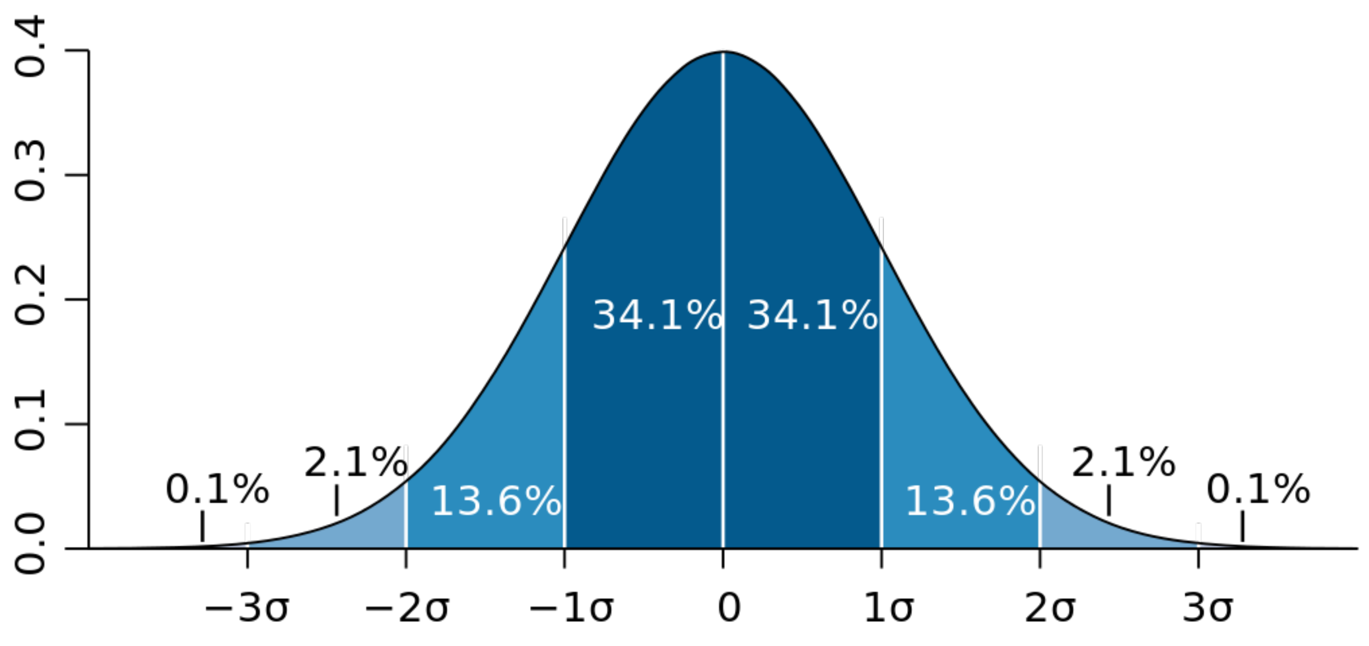
OVERVIEW

- What are GPs?
- Why are they so powerful & useful?
- What are their limitations?

PREREQUISITES

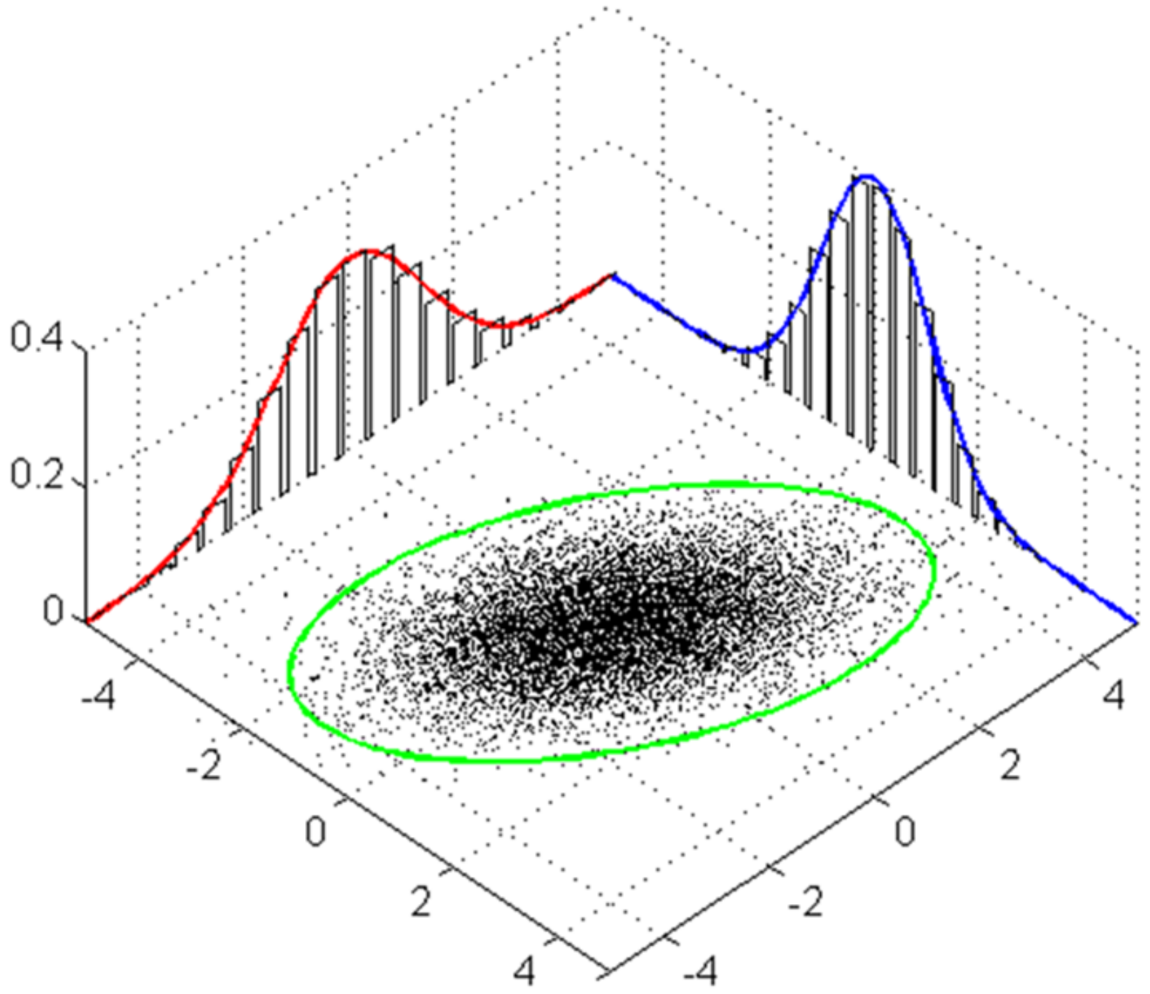
THE GAUSSIAN (NORMAL) DISTRIBUTION

Univariate: $f(x) = \dfrac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$



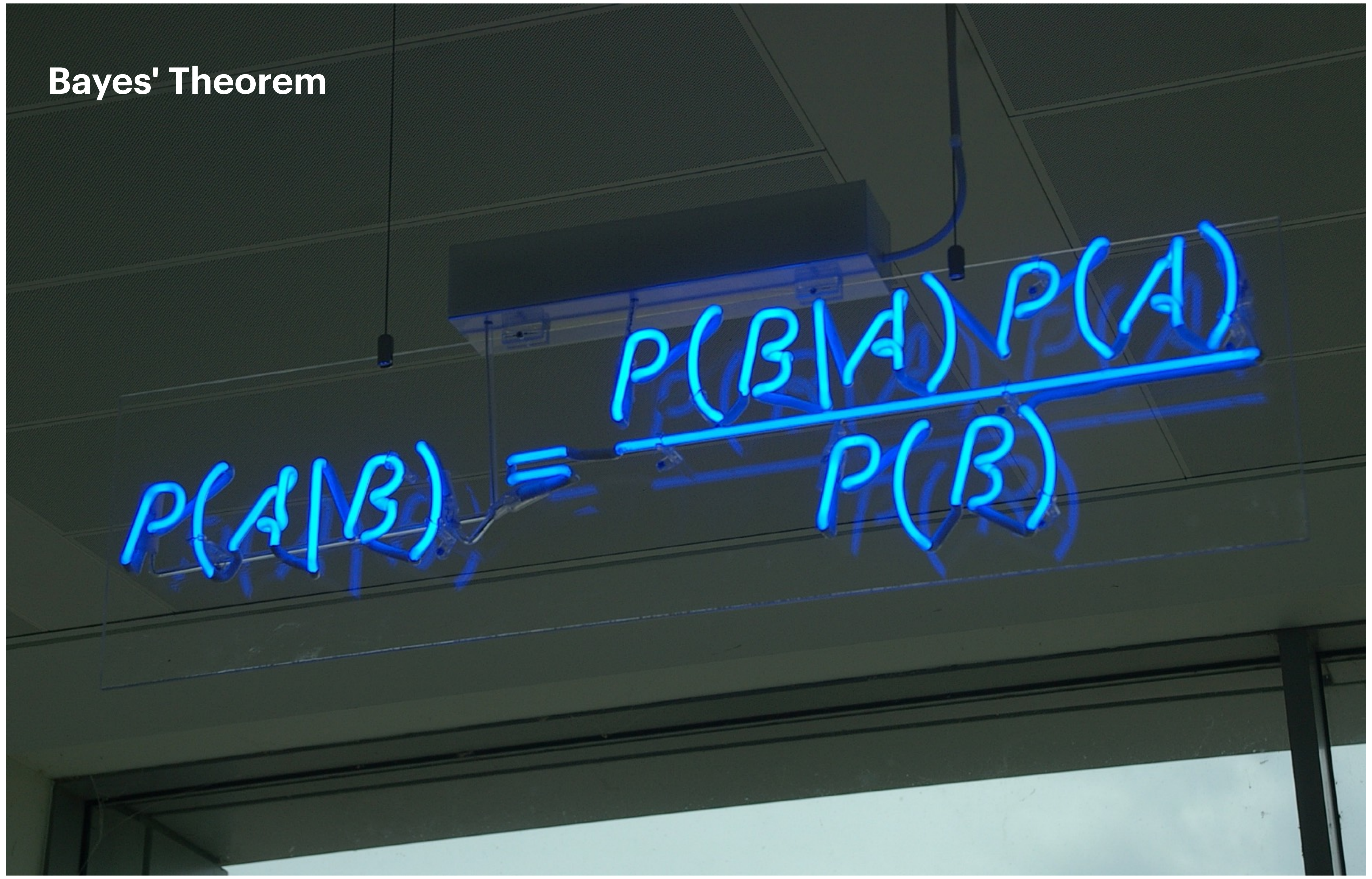
Multivariate:

$$f(\mathbf{x}) = f(x_1, \ldots, x_k) = \frac{1}{\sqrt{(2\pi)^k |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$



PREREQUISITES

BASIC BAYESIAN INFERENCE



likelihood prior

$$P(\theta\,|\,\mathscr{D}) = \frac{P(\mathscr{D}\,|\,\theta)P(\theta)}{P(\mathscr{D})}$$

posterior

evidence/marginal likelihood

Quick introductory tutorial

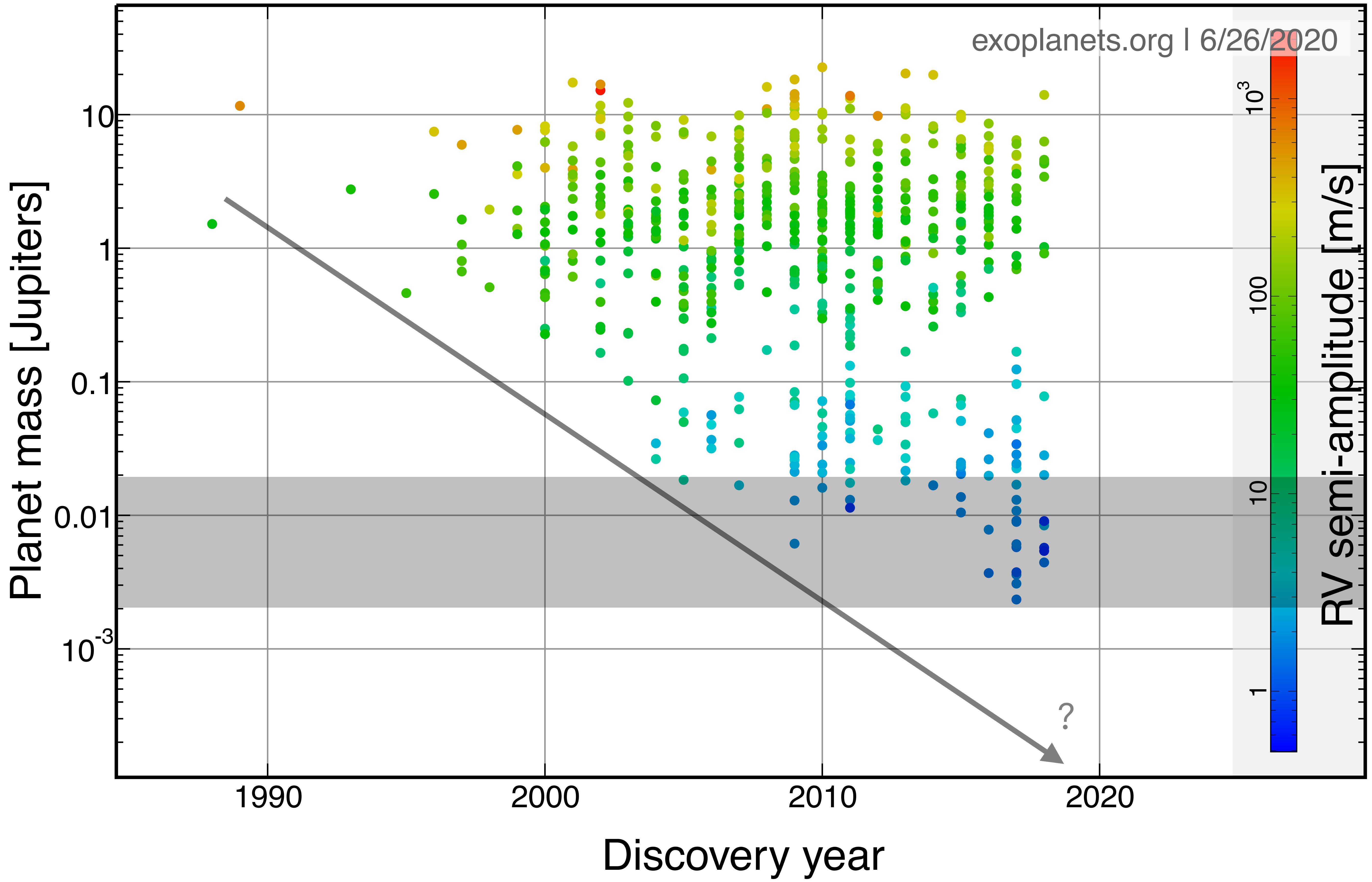
Bayesian Methods for Exoplanet Science (Parviainen, 2017; arXiv:1711.03329)

More detailed references (textbooks)

Data Analysis: A Bayesian Tutorial (Sivia & Skilling, 2006)
Bayesian Logical Data Analysis for the Physical Sciences (Gregory, 2005)

EXOPLANETS

RV DISCOVERIES vs TIME



- Where are all the RV < 1 m/s detections?
- True Earth-analogue: 9 cm/s RV signal

STARS ARE ACTIVE

Image credit NASA/SDO

STELLAR ACTIVITY

SOME SOURCES

- Signals intrinsic to **stars give rise to RV variability**
- Minutes, hours: **oscillation**, **granulation**
- Days to years: **rotationally-modulated activity**

  \+ long-term **magnetic cycles**

REALLY BAD NEWS

STELLAR ACTIVITY

STELLAR OR PLANETARY SIGNAL...?

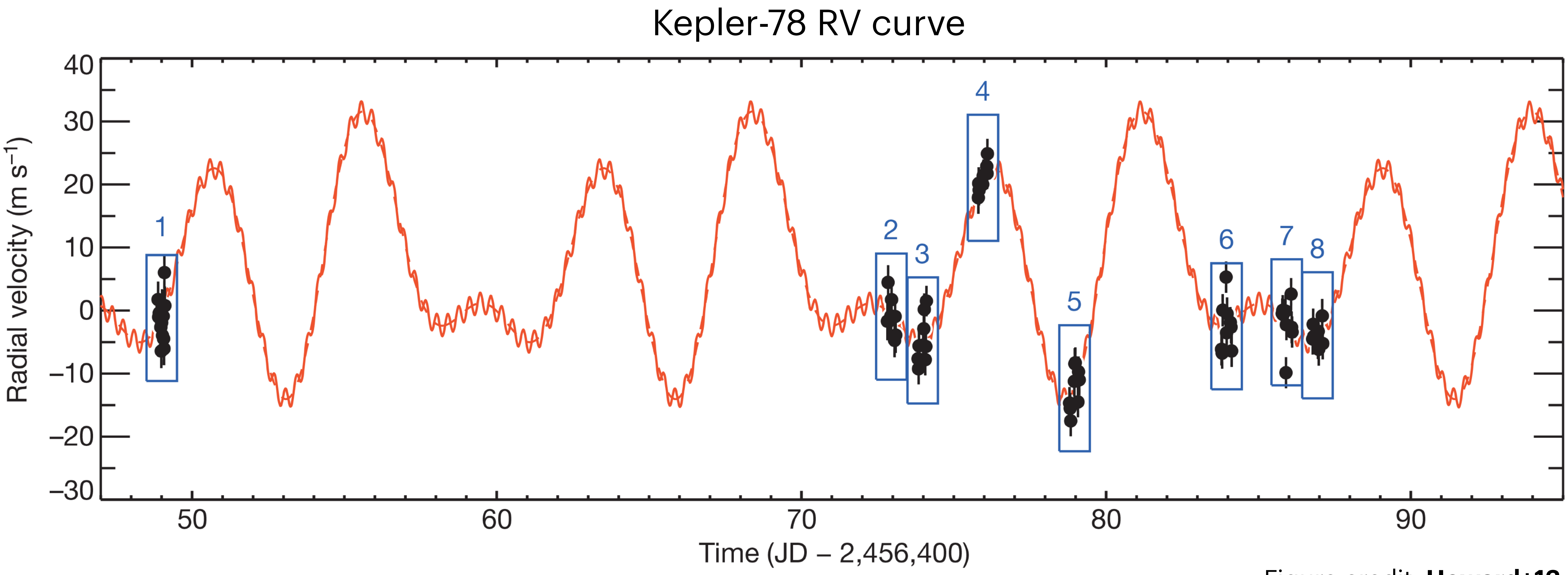
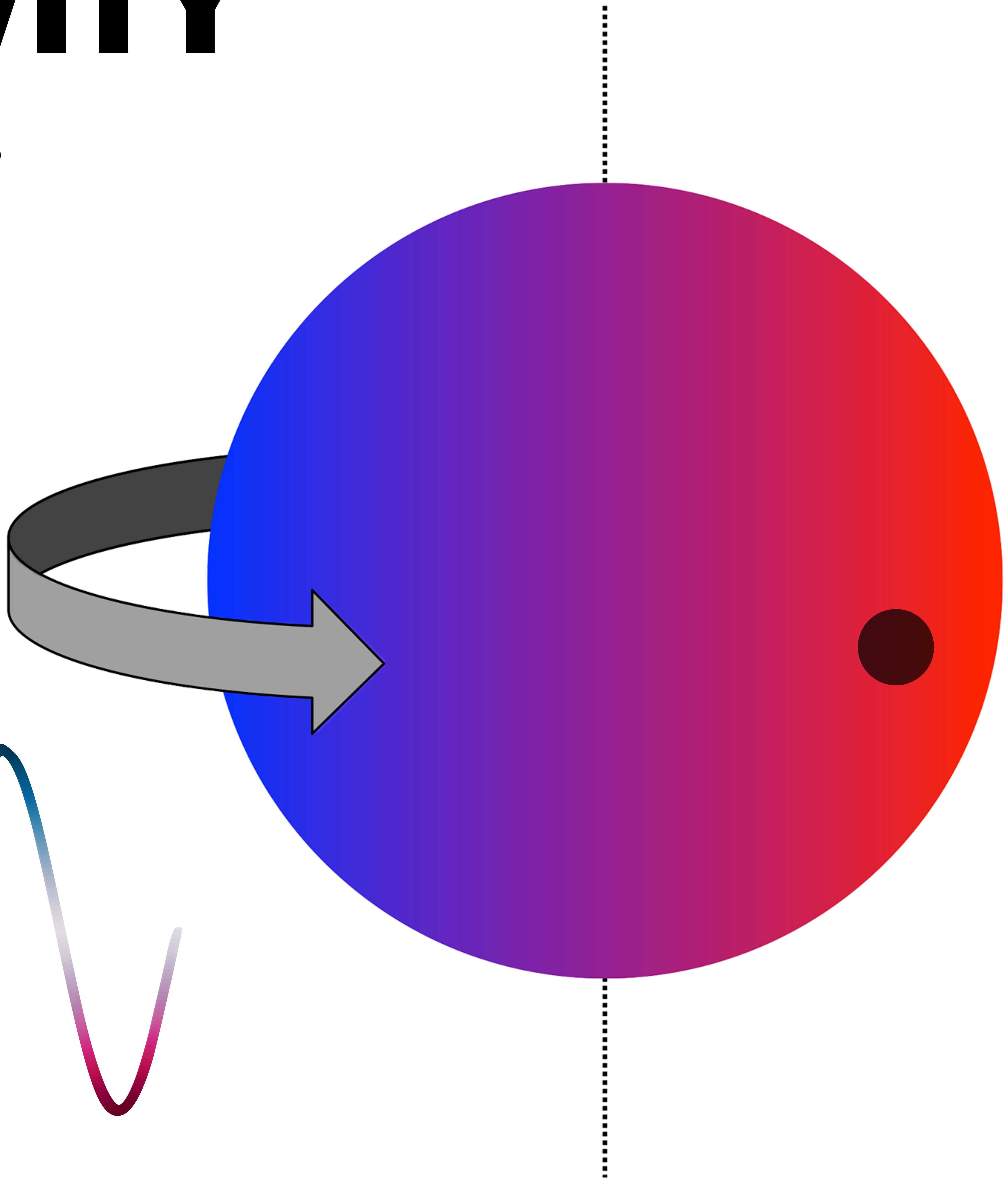


Figure credit **Howard+13**

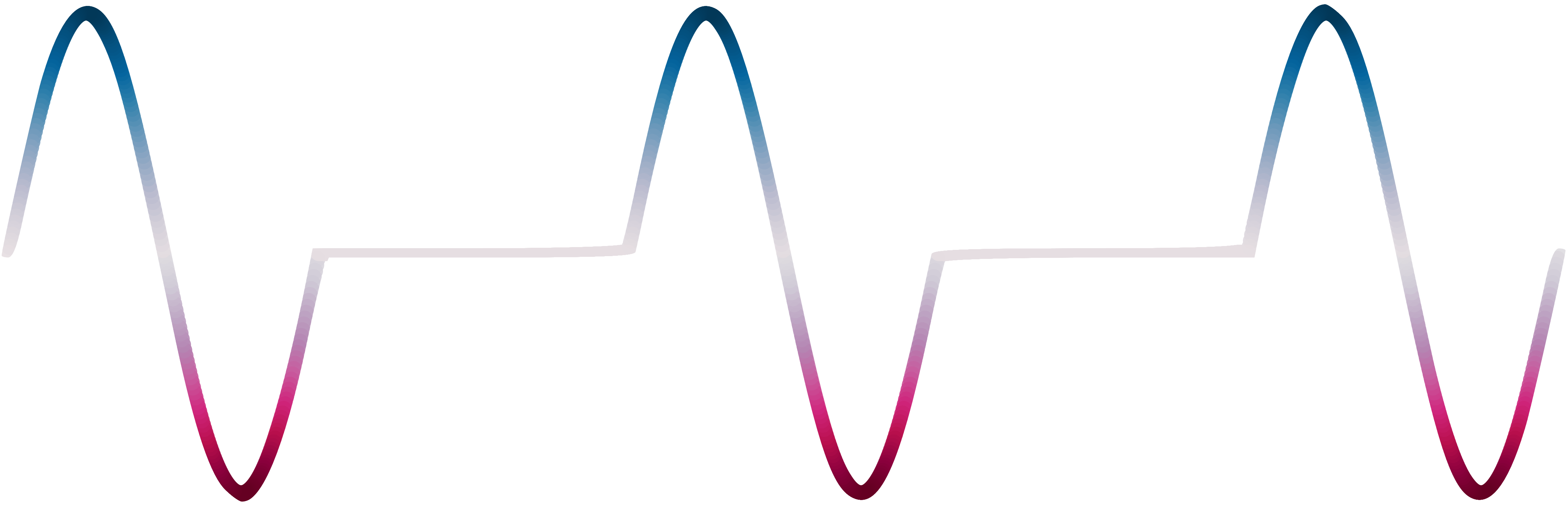
Movie credit **NASA/SDO**

STELLAR ACTIVITY

HOW ROTATIONAL ACTIVITY → RV SIGNALS



Measured stellar RV



STELLAR ACTIVITY

PROPERTIES OF ROTATIONALLY-MODULATED SIGNALS

- **Time scales similar** to those associated with **planets** (days to years)
- **Quasi-periodic** (periodic stellar rotation + evolving active regions + activity cycles)
- **Some degree of smoothness** (active regions don't change instantaneously)
- **Stochastic** (active regions seem to appear randomly)

GAUSSIAN PROCESSES

...WHAT ARE THEY? (AND WHY SHOULD YOU CARE?)

INTRODUCING GPs

WHERE NOT TO START

WIKIPEDIA
The Free Encyclopedia

Not logged in Talk Contributions Create account Log in

Article Talk — Read Edit View history — Search Wikipedia

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

Contribute
Help
Community portal
Recent changes
Upload file

Tools
What links here
Related changes
Special pages
Permanent link
Page information
Wikidata item
Cite this page

Print/export
Download as PDF
Printable version

Languages
Deutsch
Español
فارسی
Français
Italiano
日本語
Português

Gaussian process

From Wikipedia, the free encyclopedia

In probability theory and statistics, a **Gaussian process** is a stochastic process (a collection of random variables indexed by time or space), such that every finite collection of those random variables has a multivariate normal distribution, i.e. every finite linear combination of them is normally distributed. The distribution of a Gaussian process is the joint distribution of all those (infinitely many) random variables, and as such, it is a distribution over functions with a continuous domain, e.g. time or space.

A machine-learning algorithm that involves a Gaussian process uses lazy learning and a measure of the similarity between points (the *kernel function*) to predict the value for an unseen point from training data. The prediction is not just an estimate for that point, but also has uncertainty information—it is a one-dimensional Gaussian distribution.[1] For multi-output predictions, multivariate Gaussian processes [2] are used, for which the multivariate Gaussian distribution is the marginal distribution at each point.

For some kernel functions, matrix algebra can be used to calculate the predictions using the technique of kriging. When a parameterised kernel is used, optimisation software is typically used to fit a Gaussian process model.

The concept of Gaussian processes is named after Carl Friedrich Gauss because it is based on the notion of the Gaussian distribution (normal distribution). Gaussian processes can be seen as an infinite-dimensional generalization of multivariate normal distributions.

Gaussian processes are useful in statistical modelling, benefiting from properties inherited from the normal distribution. For example, if a random process is modelled as a Gaussian process, the distributions of various derived quantities can be obtained explicitly. Such quantities include the average value of the process over a range of times and the error in estimating the average using sample values at a small set of times.

Contents [show]

Definition [edit]

A time continuous stochastic process $\{X_t; t \in T\}$ is Gaussian if and only if for every finite set of indices $t_1, \ldots, t_k$ in the index set T

$$\mathbf{X}_{t_1,\ldots,t_k} = (X_{t_1}, \ldots, X_{t_k})$$

is a multivariate Gaussian random variable.[3] That is the same as saying every linear combination of $(X_{t_1}, \ldots, X_{t_k})$ has a univariate normal (or Gaussian) distribution.

Using characteristic functions of random variables, the Gaussian property can be formulated as follows: $\{X_t; t \in T\}$ is Gaussian if and only if, for every finite set of indices $t_1, \ldots, t_k$, there are real-valued $\sigma_{\ell j}$, μ_ℓ with $\sigma_{jj} > 0$ such that the following equality holds for all $s_1, s_2, \ldots, s_k \in \mathbb{R}$

$$\mathrm{E}\left(\exp\left(i \sum_{\ell=1}^{k} s_\ell\, \mathbf{X}_{t_\ell}\right)\right) = \exp\left(-\frac{1}{2} \sum_{\ell,j} \sigma_{\ell j} s_\ell s_j + i \sum_{\ell} \mu_\ell s_\ell\right).$$

where i denotes the imaginary unit such that $i^2 = -1$.

The numbers $\sigma_{\ell j}$ and μ_ℓ can be shown to be the covariances and means of the variables in the process.[4]

Variance [edit]

The variance of a Gaussian process is finite at any time t, formally[5]:p. 515

$$\mathrm{var}[X(t)] = \mathrm{E}[|X(t) - \mathrm{E}[X(t)]|^2] < \infty \quad \text{for all } t \in T.$$

INTRODUCING GPs

WHERE NOT TO START

C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X.

input: K (covariance matrix), $\mathbf{y}$ (± 1 targets)
2: $\tilde{\boldsymbol{\nu}} := \mathbf{0},\ \tilde{\boldsymbol{\tau}} := \mathbf{0},\ \Sigma := K,\ \boldsymbol{\mu} := \mathbf{0}$ — initialization and eq. (3.53)
repeat
4: **for** $i := 1$ to n **do**
$\tau_{-i} := \sigma_i^{-2} - \tilde{\tau}_i$
6: $\nu_{-i} := \sigma_i^{-2}\mu_i - \tilde{\nu}_i$ — compute approximate cavity parameters ν_{-i} and τ_{-i} using eq. (3.56)
compute the marginal moments $\hat{\mu}_i$ and $\hat{\sigma}_i^2$ — using eq. (3.58)
8: $\Delta\tilde{\tau} := \hat{\sigma}_i^{-2} - \tau_{-i} - \tilde{\tau}_i$ and $\tilde{\tau}_i := \tilde{\tau}_i + \Delta\tilde{\tau}$
$\tilde{\nu}_i := \hat{\sigma}_i^{-2}\hat{\mu}_i - \nu_{-i}$ — update site parameters $\tilde{\tau}_i$ and $\tilde{\nu}_i$ using eq. (3.59)
10: $\Sigma := \Sigma - \big((\Delta\tilde{\tau})^{-1} + \Sigma_{ii}\big)^{-1}\mathbf{s}_i\mathbf{s}_i^\top$
$\boldsymbol{\mu} := \Sigma\tilde{\boldsymbol{\nu}}$ — update Σ and $\boldsymbol{\mu}$ by eq. (3.70) and eq. (3.53). $\mathbf{s}_i$ is column i of Σ
12: **end for**
$L := \text{cholesky}(I_n + \tilde{S}^{\frac{1}{2}}K\tilde{S}^{\frac{1}{2}})$
14: $V := L^\top \backslash \tilde{S}^{\frac{1}{2}}K$
$\Sigma := K - V^\top V$ and $\boldsymbol{\mu} := \Sigma\tilde{\boldsymbol{\nu}}$ — re-compute the approximate posterior parameters Σ and $\boldsymbol{\mu}$ using eq. (3.53) and eq. (3.68)
16: **until** convergence
compute $\log Z_{\text{EP}}$ — using eq. (3.65), (3.73) and (3.74) and the existing L
18: **return**: $\tilde{\boldsymbol{\nu}},\ \tilde{\boldsymbol{\tau}}$ (natural site param.), $\log Z_{\text{EP}}$ (approx. log marg. likelihood)

Algorithm 3.5: Expectation Propagation for binary classification. The targets $\mathbf{y}$ are used only in line 7. In lines 13-15 the parameters of the approximate posterior are re-computed (although they already exist); this is done because of the large number of rank-one updates in line 10 which would eventually cause loss of numerical precision in Σ. The computational complexity is dominated by the rank-one updates in line 10, which takes $\mathcal{O}(n^2)$ per variable, i.e. $\mathcal{O}(n^3)$ for an entire sweep over all variables. Similarly re-computing Σ in lines 13-15 is $\mathcal{O}(n^3)$.

the eigenvalues of B are bounded below by one. The parameters of the Gaussian approximate posterior from eq. (3.53) are computed as

$$\Sigma = (K^{-1} + \tilde{S})^{-1} = K - K(K + \tilde{S}^{-1})^{-1}K = K - K\tilde{S}^{\frac{1}{2}}B^{-1}\tilde{S}^{\frac{1}{2}}K. \quad (3.68)$$

After updating the parameters of a site, we need to update the approximate posterior eq. (3.53) taking the new site parameters into account. For the inverse covariance matrix of the approximate posterior we have from eq. (3.53)

$$\Sigma^{-1} = K^{-1} + \tilde{S}, \quad \text{and thus} \quad \Sigma_{\text{new}}^{-1} = K^{-1} + \tilde{S}_{\text{old}} + (\tilde{\tau}_i^{\text{new}} - \tilde{\tau}_i^{\text{old}})\mathbf{e}_i\mathbf{e}_i^\top, \quad (3.69)$$

where $\mathbf{e}_i$ is a unit vector in direction i, and we have used that $\tilde{S} = \text{diag}(\tilde{\boldsymbol{\tau}})$. Using the matrix inversion lemma eq. (A.9), on eq. (3.69) we obtain the new Σ

$$\Sigma^{\text{new}} = \Sigma^{\text{old}} - \frac{\tilde{\tau}_i^{\text{new}} - \tilde{\tau}_i^{\text{old}}}{1 + (\tilde{\tau}_i^{\text{new}} - \tilde{\tau}_i^{\text{old}})\Sigma_{ii}^{\text{old}}}\mathbf{s}_i\mathbf{s}_i^\top, \quad (3.70)$$

in time $\mathcal{O}(n^2)$, where $\mathbf{s}_i$ is the i'th column of Σ^{old}. The posterior mean is then calculated from eq. (3.53).

In the EP algorithm each site is updated in turn, and several passes over all sites are required. Pseudocode for the EP-GPC algorithm is given in Algorithm

C. E. Rasmussen & C. K. I. Williams, Gaussian Processes for Machine Learning, the MIT Press, 2006, ISBN 026218253X. © 2006 Massachusetts Institute of Technology. www.GaussianProcess.org/gpml

input: $\tilde{\boldsymbol{\nu}},\ \tilde{\boldsymbol{\tau}}$ (natural site param.), X (inputs), $\mathbf{y}$ (± 1 targets), k (covariance function), $\mathbf{x}_*$ test input
2: $L := \text{cholesky}(I_n + \tilde{S}^{\frac{1}{2}}K\tilde{S}^{\frac{1}{2}})$ — $B = I_n + \tilde{S}^{\frac{1}{2}}K\tilde{S}^{\frac{1}{2}}$
$\mathbf{z} := \tilde{S}^{\frac{1}{2}}L^\top \backslash (L \backslash (\tilde{S}^{\frac{1}{2}}K\tilde{\boldsymbol{\nu}}))$
4: $\bar{f}_* := \mathbf{k}(\mathbf{x}_*)^\top(\tilde{\boldsymbol{\nu}} - \mathbf{z})$ — eq. (3.60) using eq. (3.71)
$\mathbf{v} := L \backslash \big(\tilde{S}^{\frac{1}{2}}\mathbf{k}(\mathbf{x}_*)\big)$
6: $\mathbb{V}[f_*] := k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{v}^\top\mathbf{v}$ — eq. (3.61) using eq. (3.72)
$\bar{\pi}_* := \Phi(\bar{f}_*/\sqrt{1 + \mathbb{V}[f_*]})$ — eq. (3.63)
8: **return**: $\bar{\pi}_*$ (predictive class probability (for class 1))

Algorithm 3.6: Predictions for expectation propagation. The natural site parameters $\tilde{\boldsymbol{\nu}}$ and $\tilde{\boldsymbol{\tau}}$ of the posterior (which can be computed using algorithm 3.5) are input. For multiple test inputs lines 4-7 are applied to each test input. Computational complexity is $n^3/6 + n^2$ operations once (line 2 and 3) plus n^2 operations per test case (line 5), although the Cholesky decomposition in line 2 could be avoided by storing it in Algorithm 3.5. Note the close similarity to Algorithm 3.2 on page 47.

3.5. There is no formal guarantee of convergence, but several authors have reported that EP for Gaussian process models works relatively well.[16]

For the predictive distribution, we get the mean from eq. (3.60) which is evaluated using

$$\begin{aligned}\mathbb{E}_q[f_*|X,\mathbf{y},\mathbf{x}_*] &= \mathbf{k}_*^\top(K + \tilde{S}^{-1})^{-1}\tilde{S}^{-1}\tilde{\boldsymbol{\nu}} = \mathbf{k}_*^\top\big(I - (K + \tilde{S}^{-1})^{-1}K\big)\tilde{\boldsymbol{\nu}} \\ &= \mathbf{k}_*^\top(I - \tilde{S}^{\frac{1}{2}}B^{-1}\tilde{S}^{\frac{1}{2}}K)\tilde{\boldsymbol{\nu}},\end{aligned} \quad (3.71)$$

and the predictive variance from eq. (3.61) similarly by

$$\begin{aligned}\mathbb{V}_q[f_*|X,\mathbf{y},\mathbf{x}_*] &= k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top(K + \tilde{S}^{-1})^{-1}\mathbf{k}_* \\ &= k(\mathbf{x}_*, \mathbf{x}_*) - \mathbf{k}_*^\top\tilde{S}^{\frac{1}{2}}B^{-1}\tilde{S}^{\frac{1}{2}}\mathbf{k}_*.\end{aligned} \quad (3.72)$$

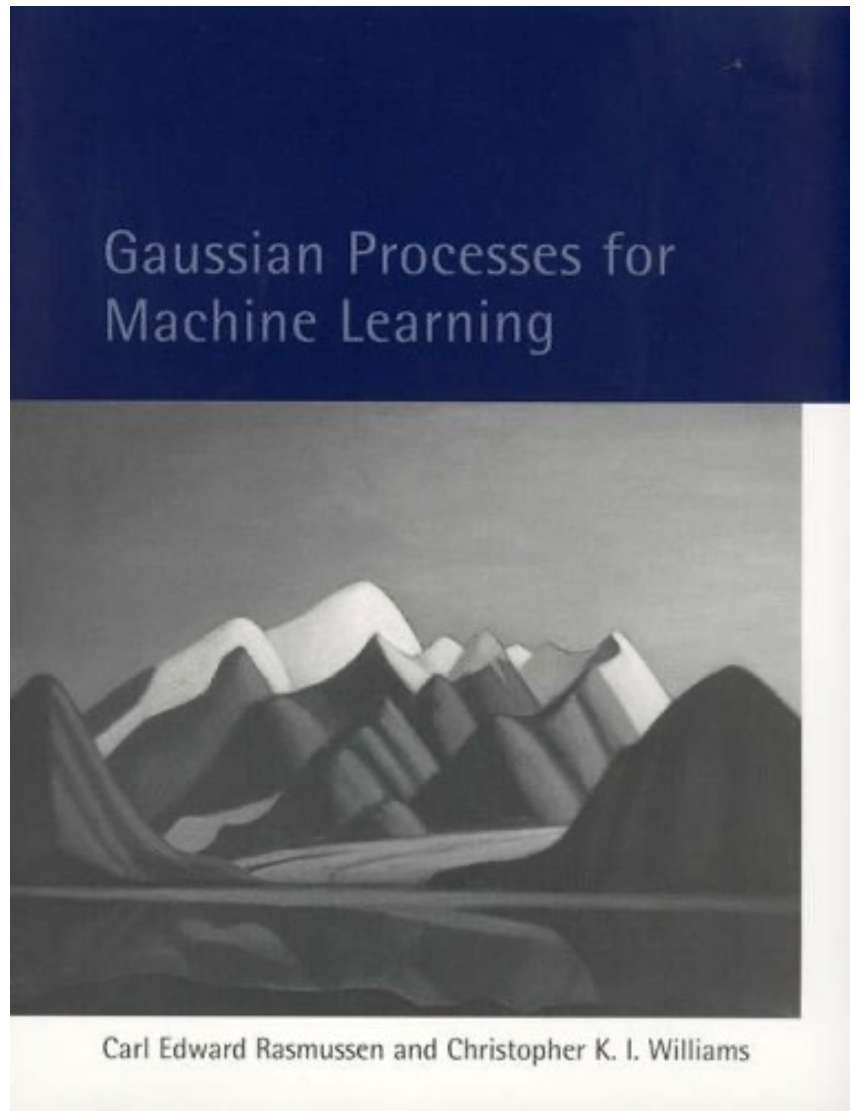
Pseudocode for making predictions using EP is given in Algorithm 3.6.

Finally, we need to evaluate the approximate log marginal likelihood from eq. (3.65). There are several terms which need careful consideration, principally due to the fact the $\tilde{\tau}_i$ values may be arbitrarily small (and cannot safely be inverted). We start with the fourth and first terms of eq. (3.65)

$$\begin{aligned}\tfrac{1}{2}\log|T^{-1}+\tilde{S}^{-1}| - \tfrac{1}{2}\log|K+\tilde{\Sigma}| &= \tfrac{1}{2}\log|\tilde{S}^{-1}(I+\tilde{S}T^{-1})| - \tfrac{1}{2}\log|\tilde{S}^{-1}B| \\ &= \tfrac{1}{2}\sum_i \log(1+\tilde{\tau}_i\tau_{-i}^{-1}) - \sum_i \log L_{ii},\end{aligned} \quad (3.73)$$

where T is a diagonal matrix of cavity precisions $T_{ii} = \tau_{-i} = \sigma_{-i}^{-2}$ and L is the Cholesky factorization of B. In eq. (3.73) we have factored out the matrix $\tilde{S}^{-1}$ from both determinants, and the terms cancel. Continuing with the part of the

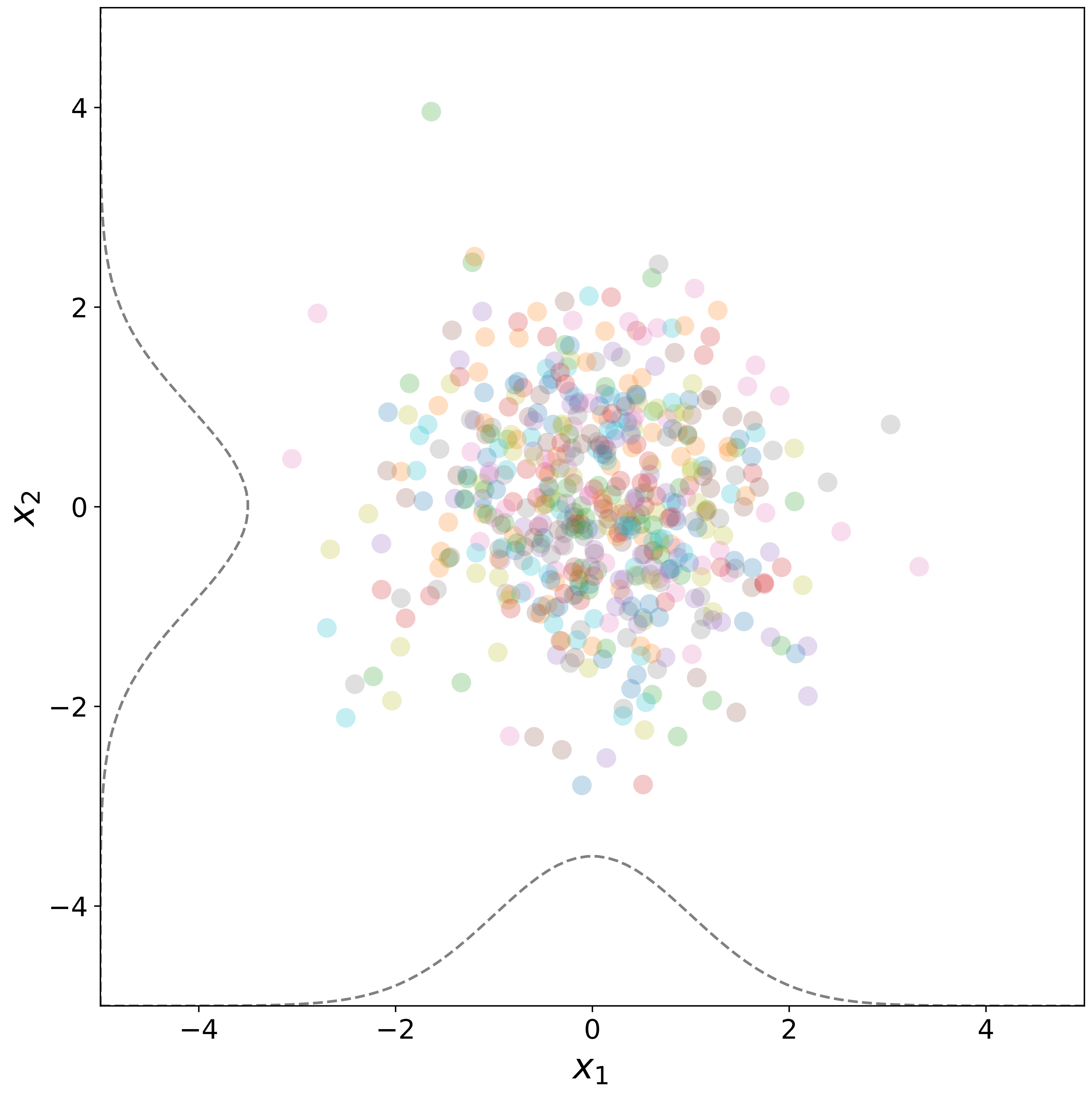
[16] It has been conjectured (but not proven) by L. Csató (personal communication) that EP is guaranteed to converge if the likelihood is log concave.



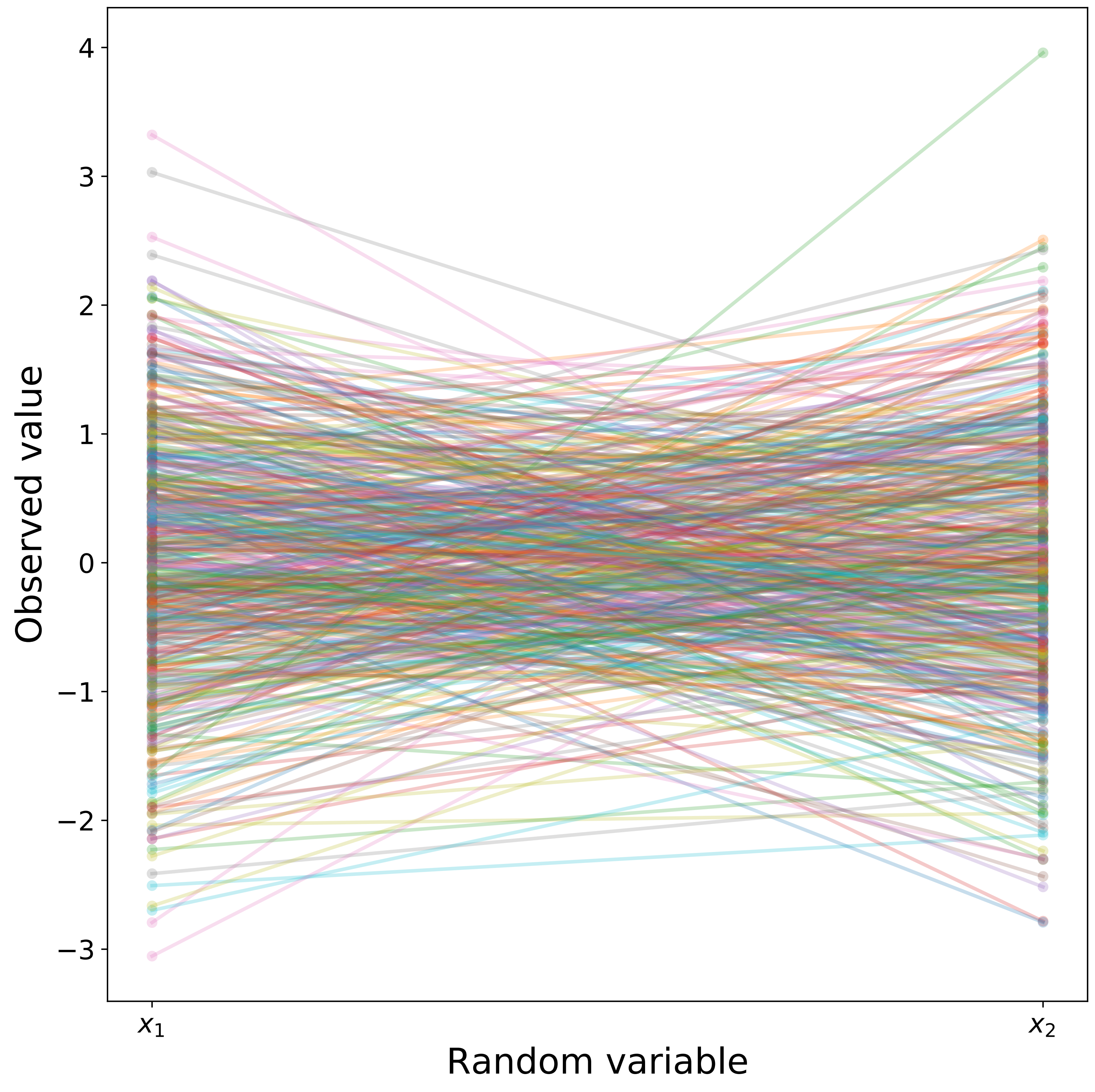
Rasmussen & Williams: very maths-heavy, and not ideal for beginners (but a brilliant & definitive GP reference nonetheless)

INTRODUCING GPs

BIVARIATE GAUSSIAN: TWO REPRESENTATIONS

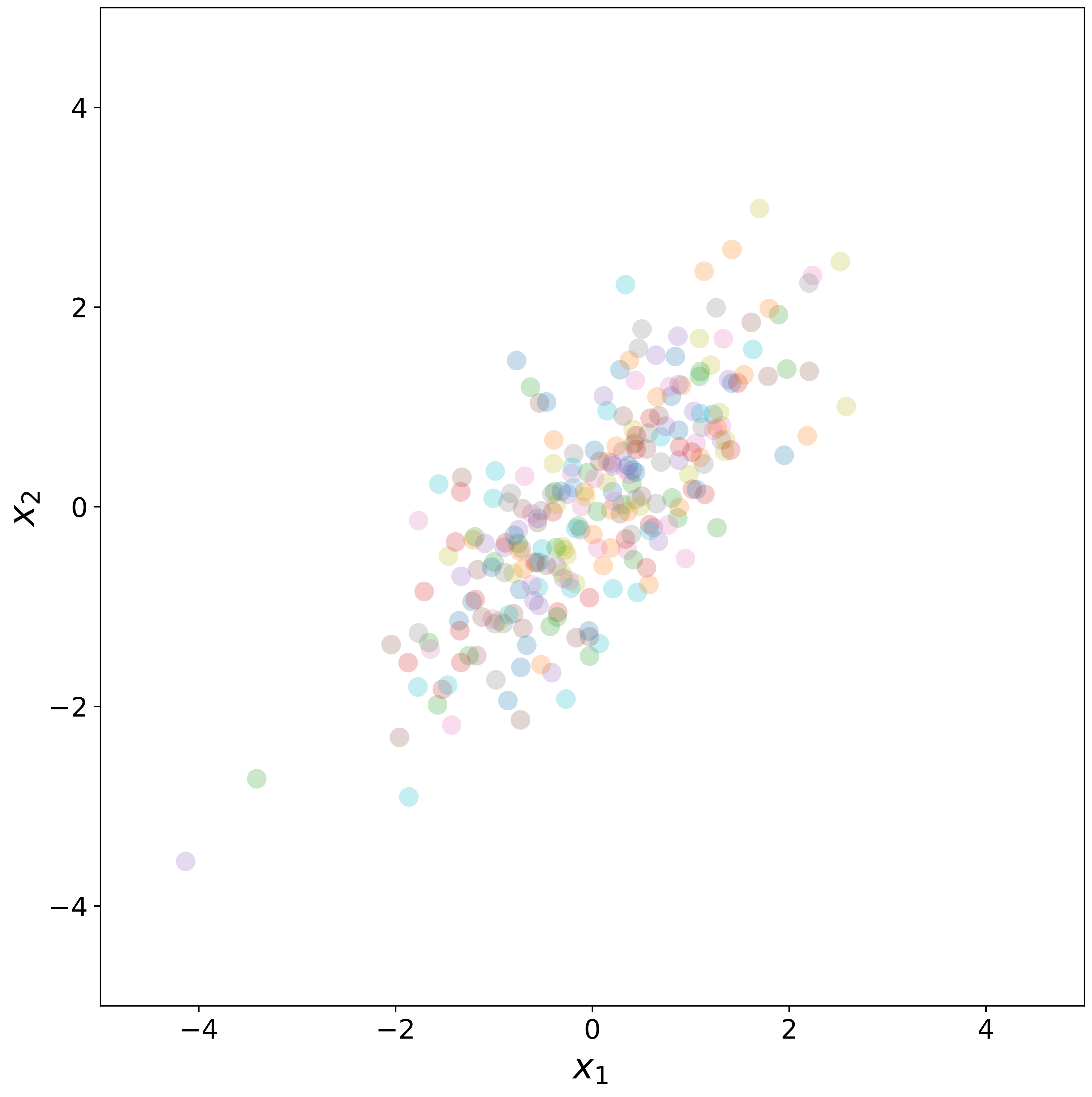


equivalent

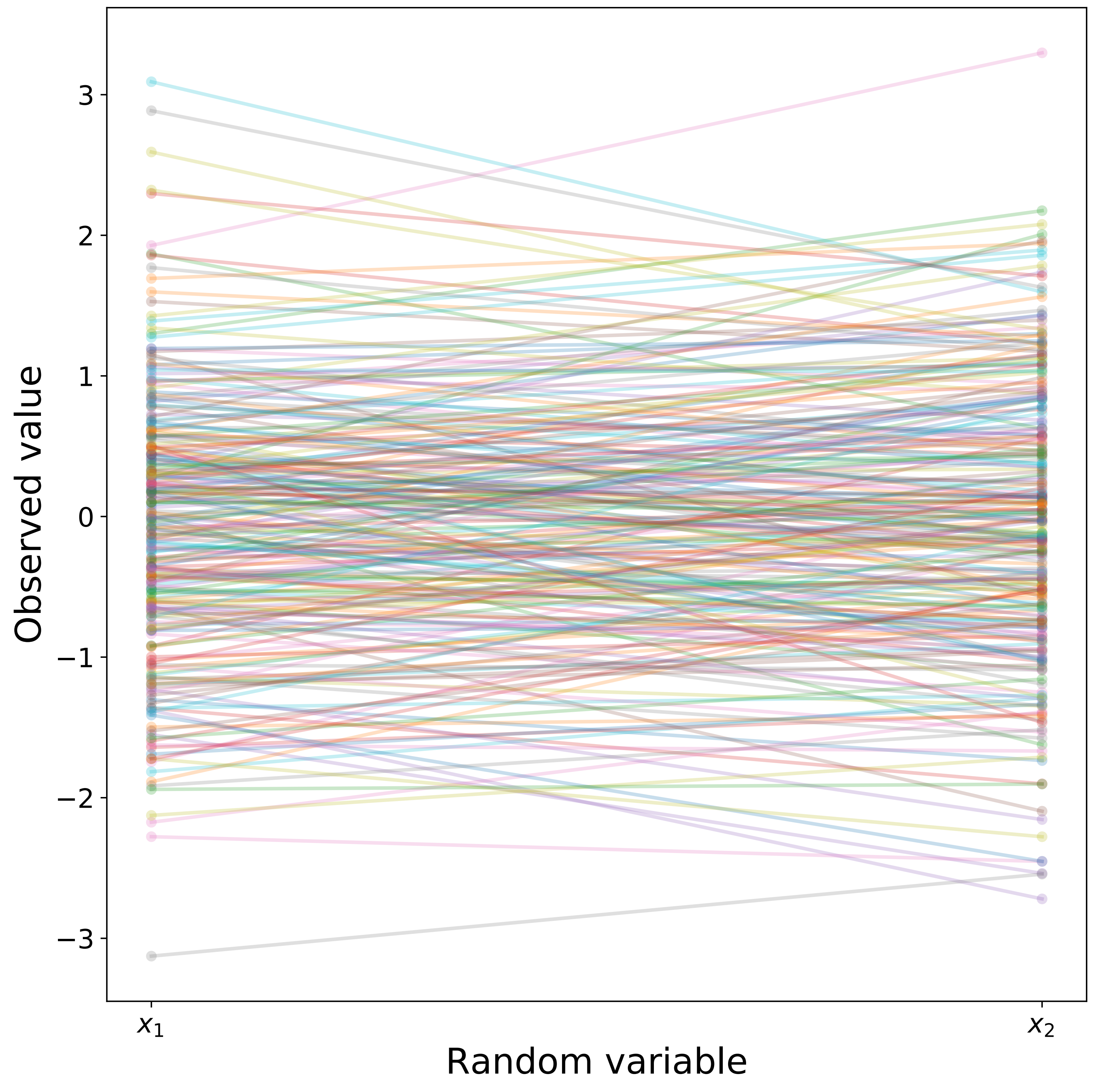


INTRODUCING GPs

BIVARIATE GAUSSIAN: TWO REPRESENTATIONS

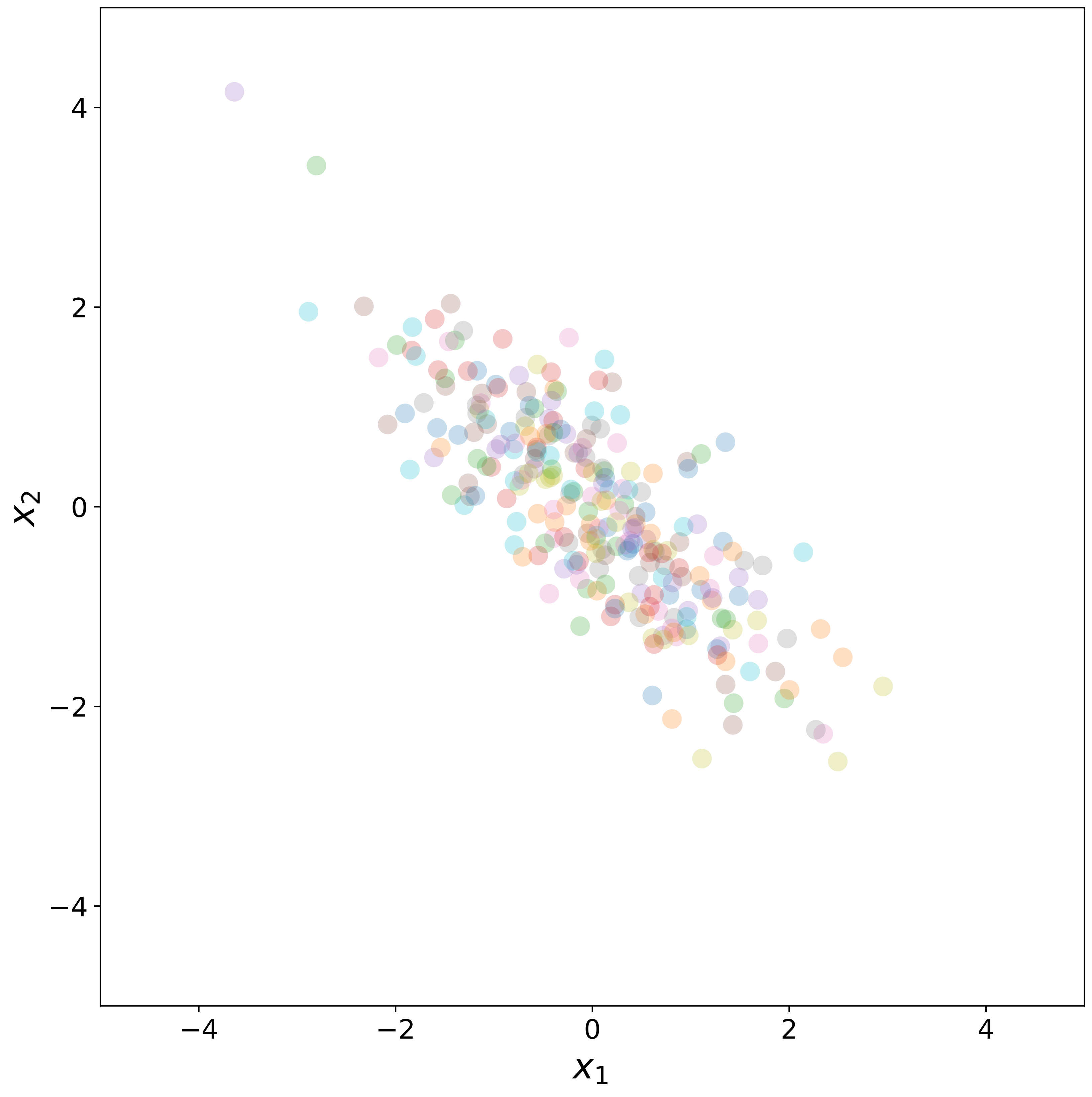


equivalent

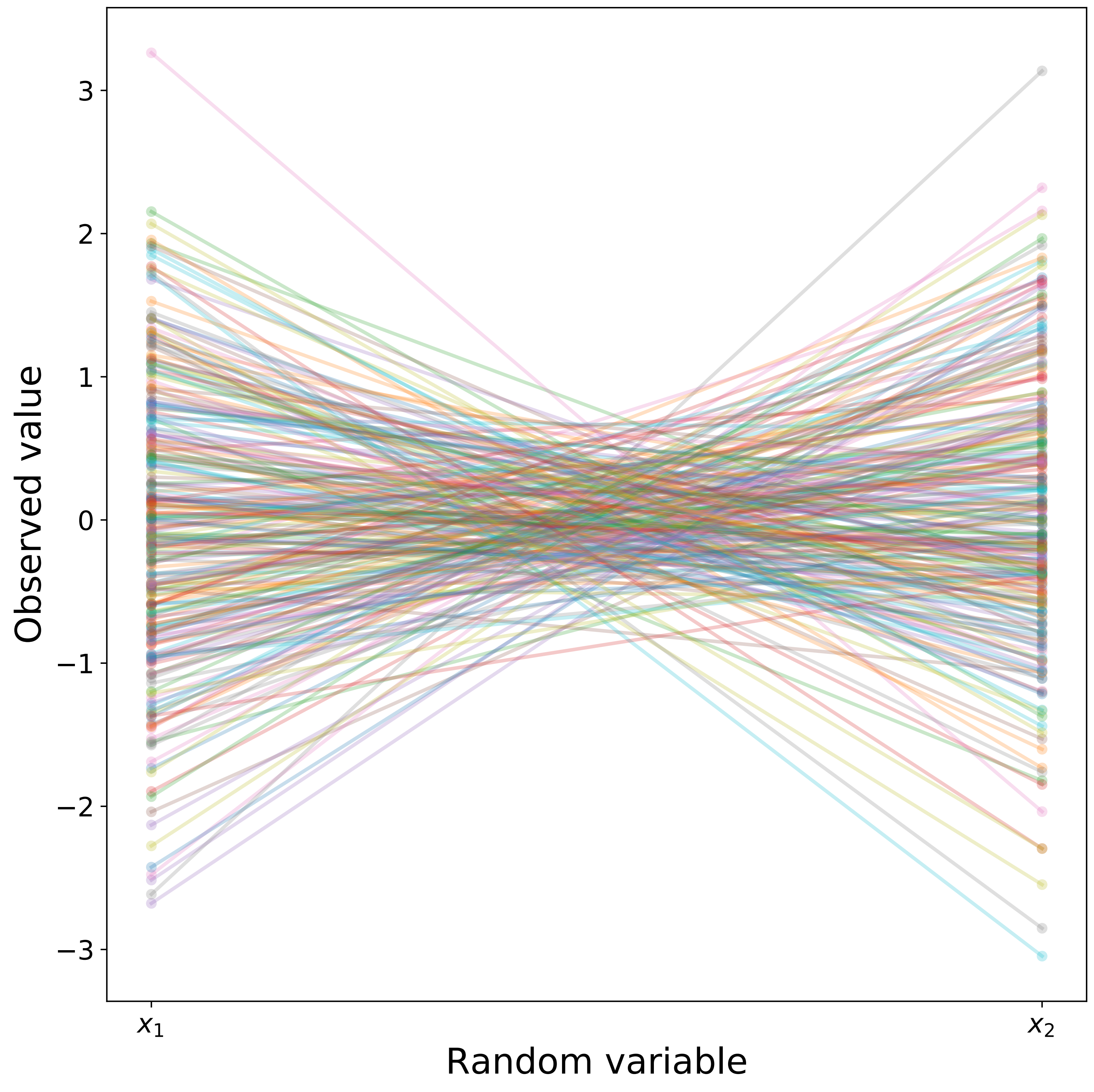


INTRODUCING GPs

BIVARIATE GAUSSIAN: TWO REPRESENTATIONS



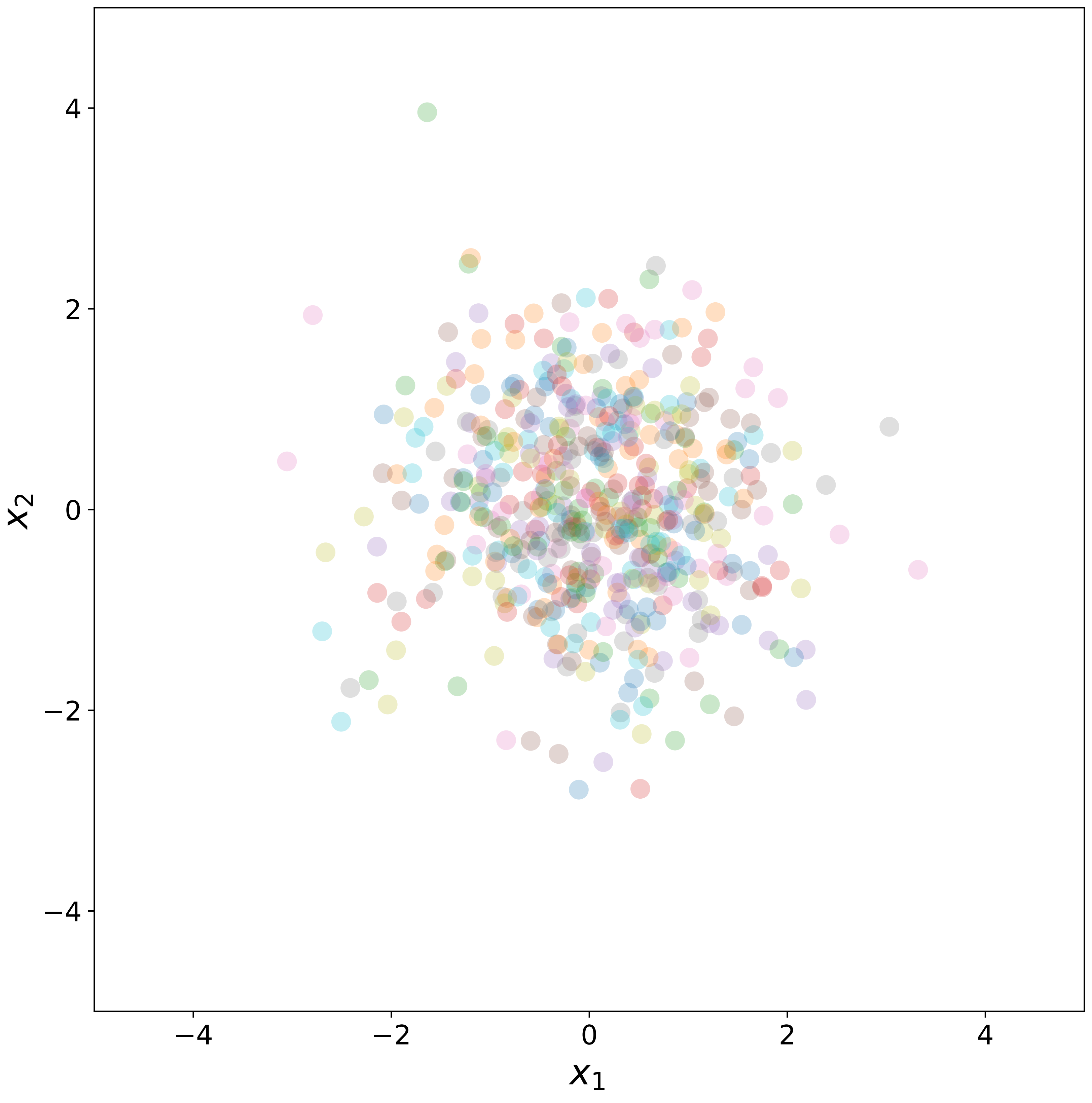
equivalent



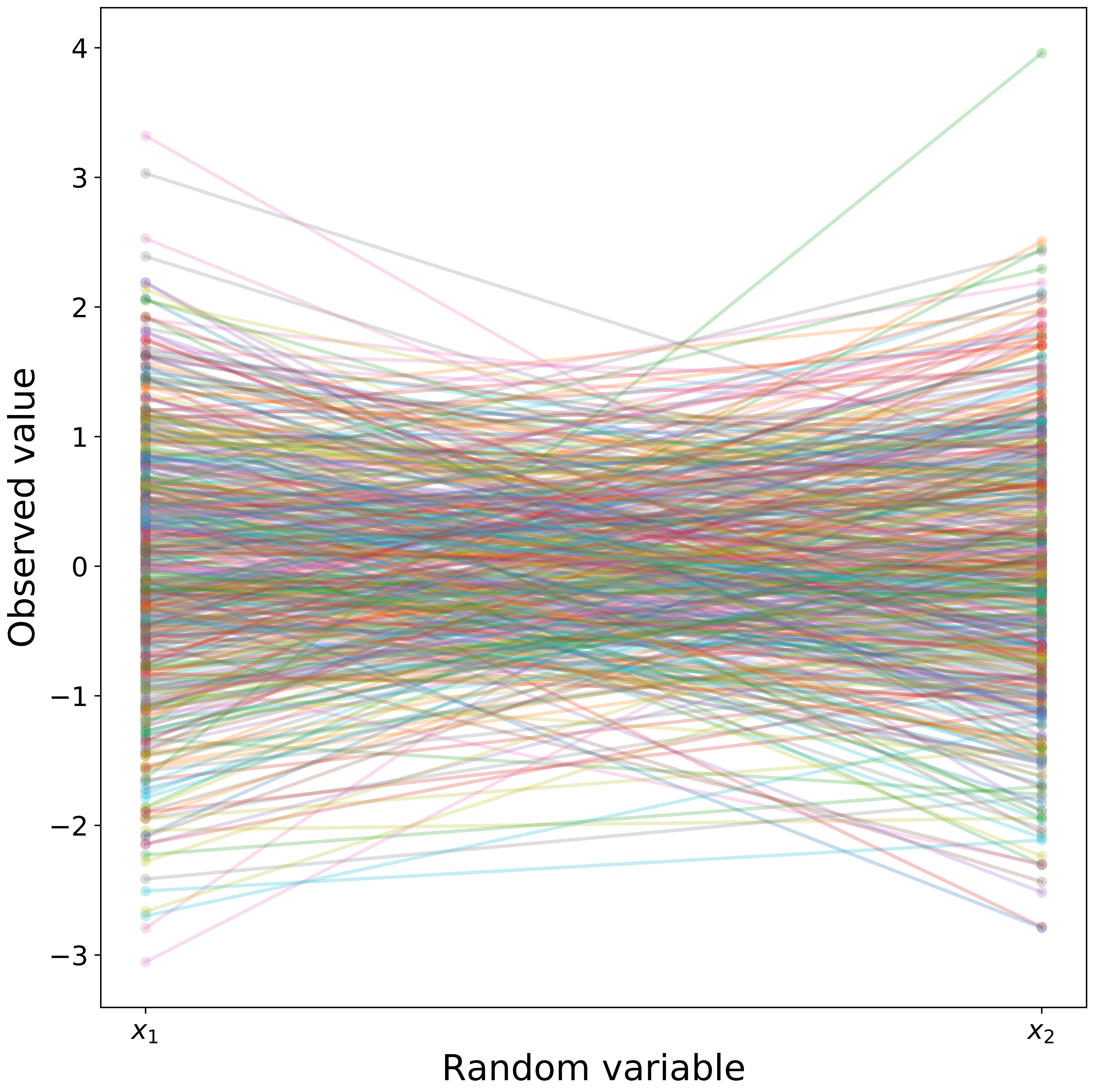
INTRODUCING GPs

BIVARIATE GAUSSIAN: TWO REPRESENTATIONS

$$k(x_1, x_2) = 0$$



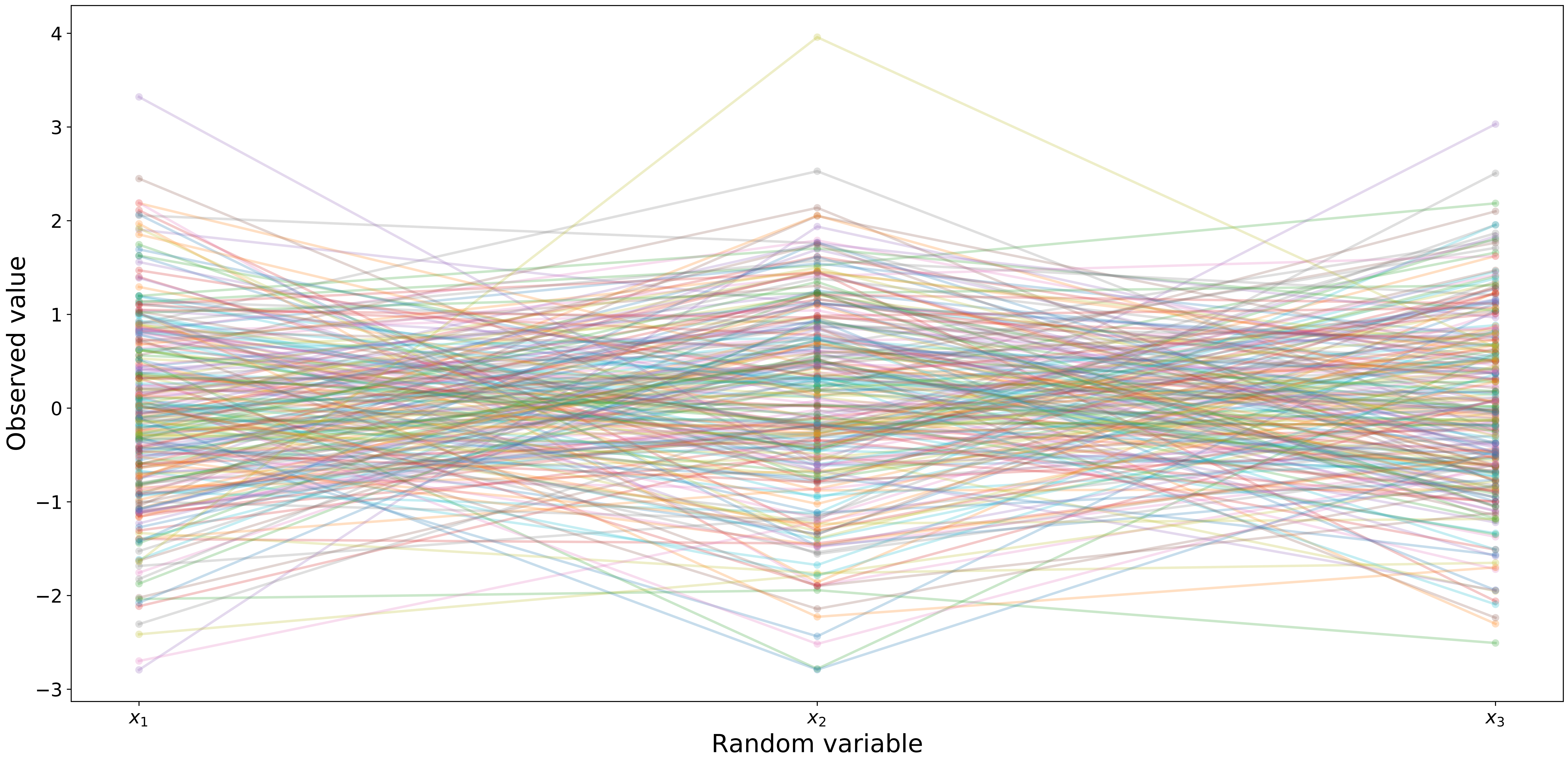
equivalent



INTRODUCING GPs

MOVING TO THREE DIMENSIONS...

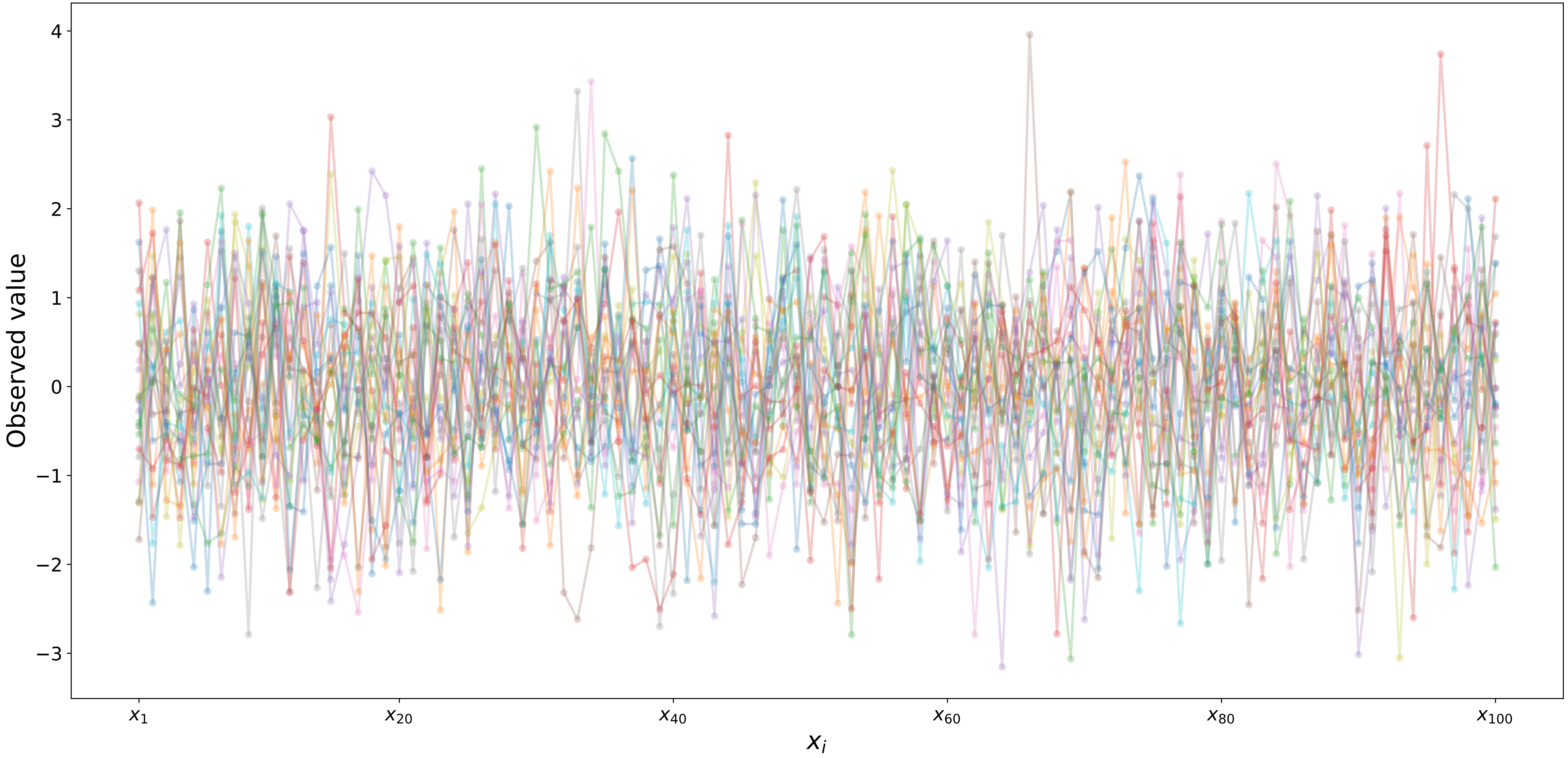
$$k(x_i, x_j) = 0, \forall\ i \neq j$$



INTRODUCING GPs

NOW A HUNDRED DIMENSIONS...

$$k(x_i, x_j) = 0, \forall\ i \neq j$$

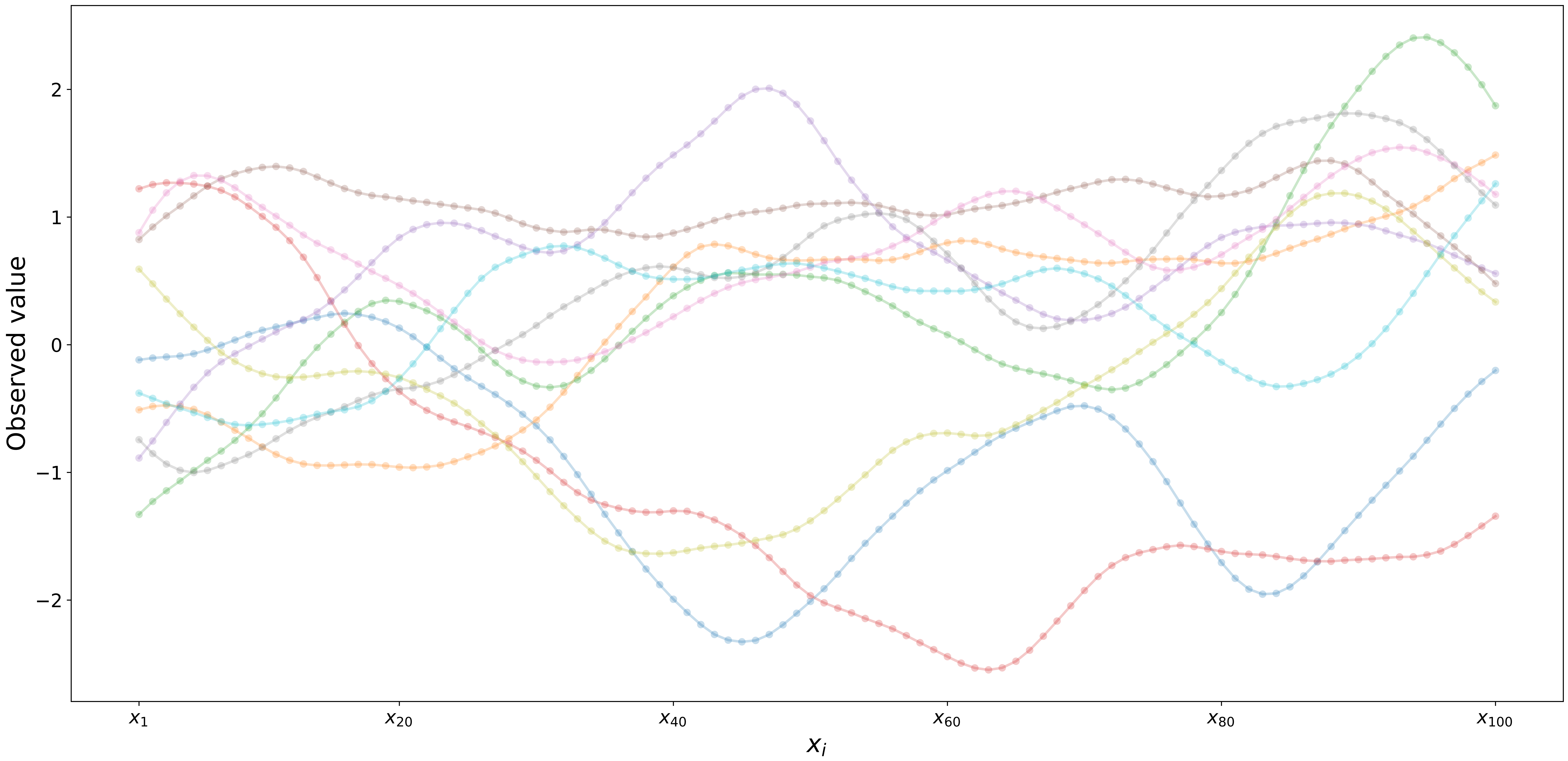


INTRODUCING GPs

$$\cancel{k(x_i, x_j) = 0, \forall\, i \neq j}$$

$$k(x_i, x_j) = e^{-|x_i - x_j|^2}$$

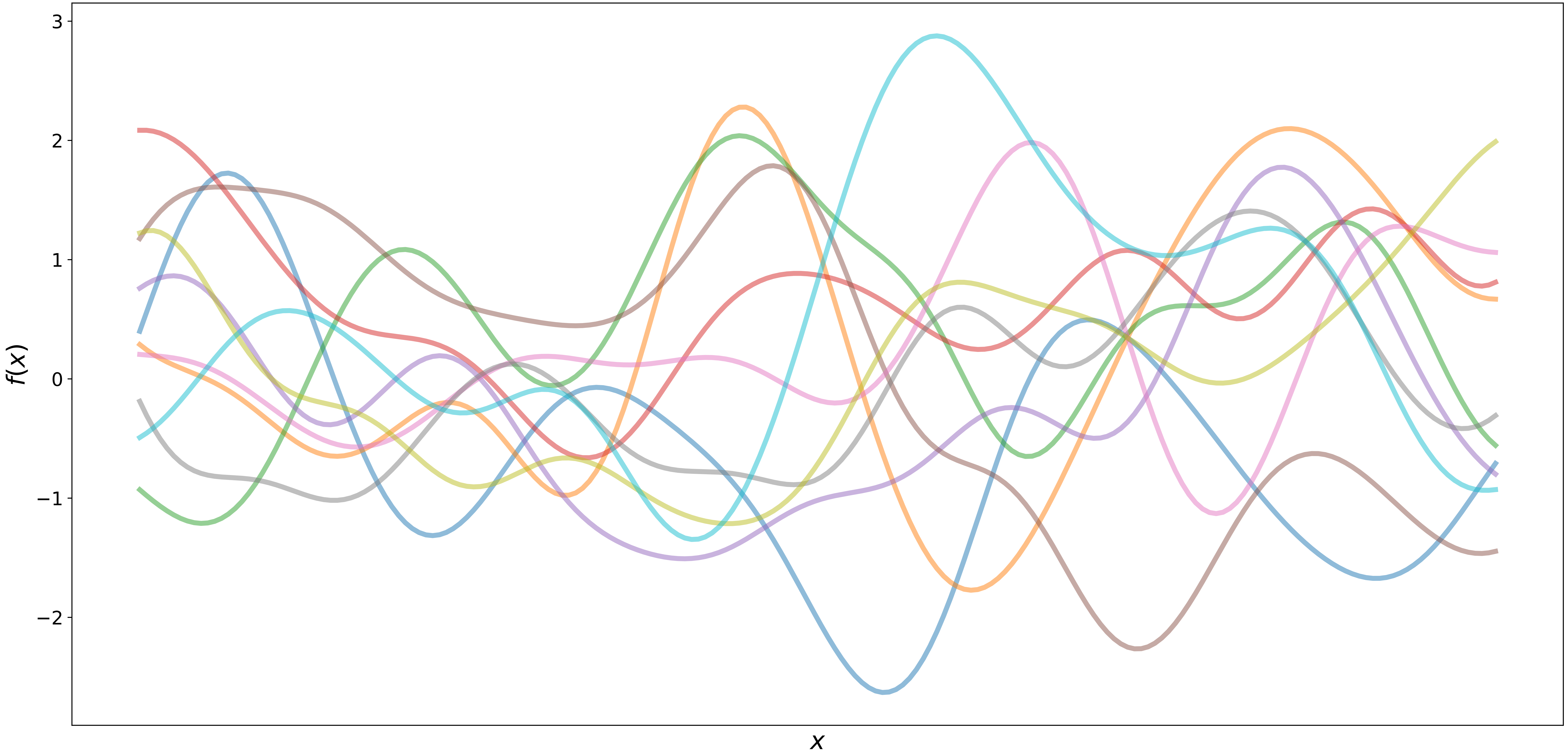
...AND NOW WITH NON-TRIVIAL COVARIANCE!



INTRODUCING GPs

...AND IN INFINITELY MANY DIMENSIONS!

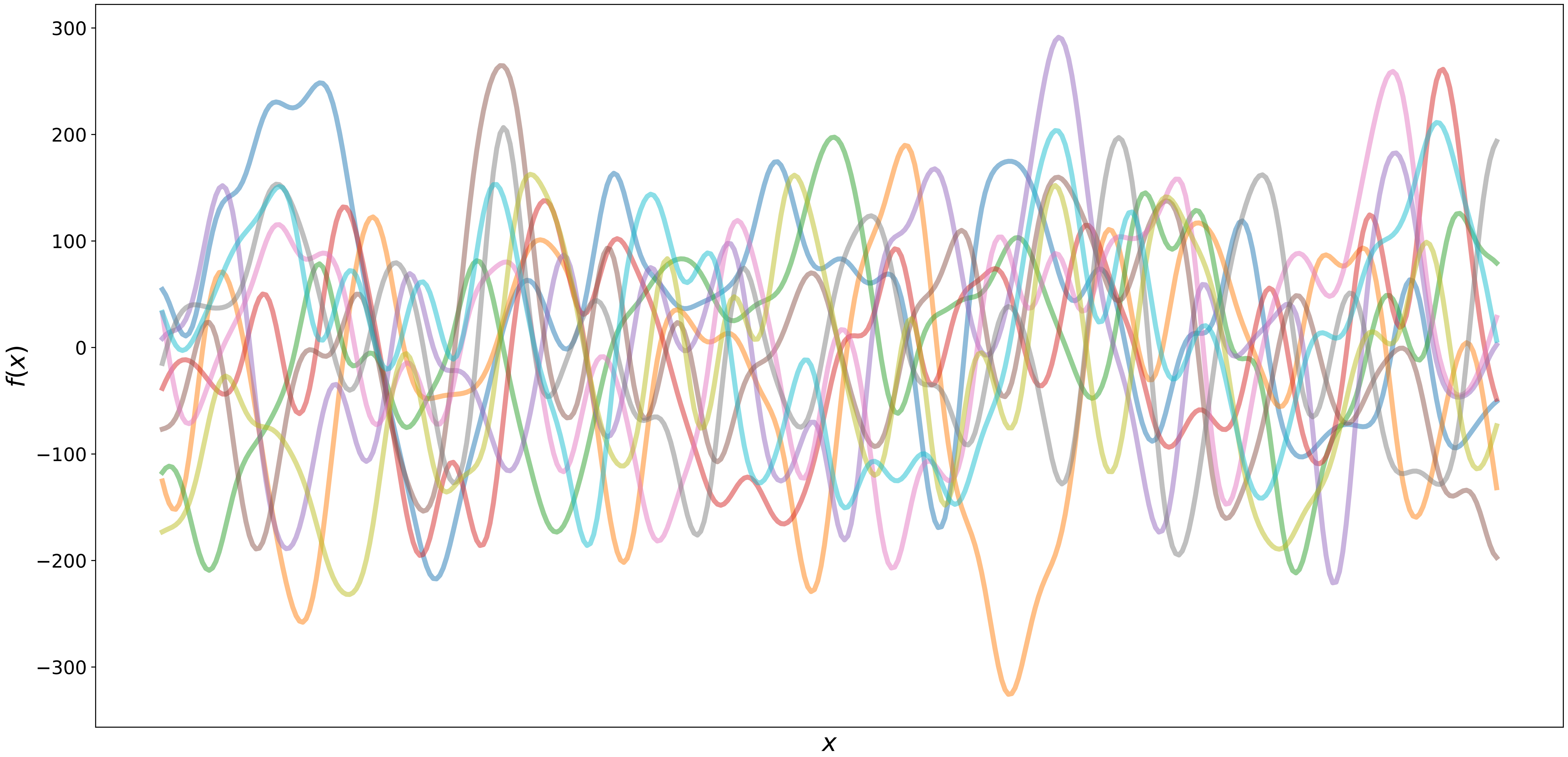
$$k(x_i, x_j) = e^{-|x_i - x_j|^2}$$



INTRODUCING GPs

TWEAKING THE COVARIANCE

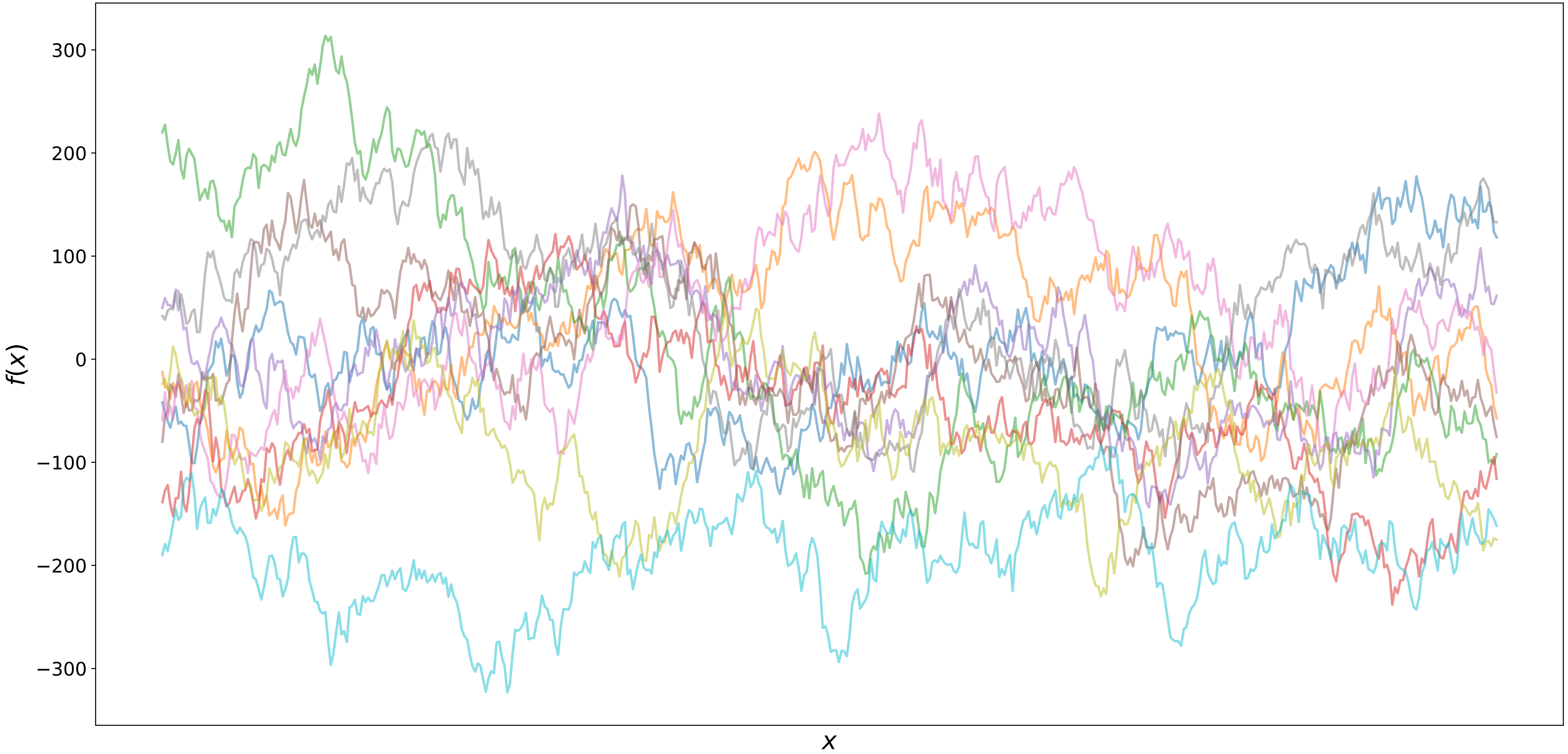
$$k(x_i, x_j) = 100e^{-2|x_i - x_j|^2}$$



INTRODUCING GPs

BROWNIAN MOTION COVARIANCE

$$k(x_i, x_j) = 100e^{-10|x_i - x_j|}$$



a **GP** is the **infinite-dimensional** version of a **Gaussian distribution**

INTRODUCING GPs

1D Gaussian	→	multivariate Gaussian	→	Gaussian process
μ		$\boldsymbol{\mu}$		$\mu(x)$
σ		$\boldsymbol{\Sigma}$		$k(x_i, x_j)$
PDF over scalars		PDF over vectors		PDF over functions

INTRODUCING GPs

THE REMARKABLE PROPERTIES OF GAUSSIANS

- The **marginalisation property** allows us to compute marginals & conditionals for arbitrary, finite subsets of variables
- Gaussian prior + likelihood → posterior that is also a GP (**conjugacy**)
- So, in practice: **GP prior + data → GP posterior distribution** that can be evaluated **analytically**
- We can use GPs to **learn unknown functions (+ error bars) directly from data!**

INTRODUCING GPs

REGRESSION WITH GPs - STRAIGHTFORWARD

If $y_i = f(t_i) + \epsilon_i$, and $f \sim \mathcal{GP}(\boldsymbol{\mu}, \mathbf{K})$

$\boldsymbol{\mu} = \mu(\boldsymbol{t}; \phi)$ $\longrightarrow$ deterministic, easy-to-model stuff (e.g. planets)

$\mathbf{K}_{ij} = k(t_i, t_j; \theta)$ $\longrightarrow$ stochastic signals/stuff we can't parametrise (e.g. stellar activity)

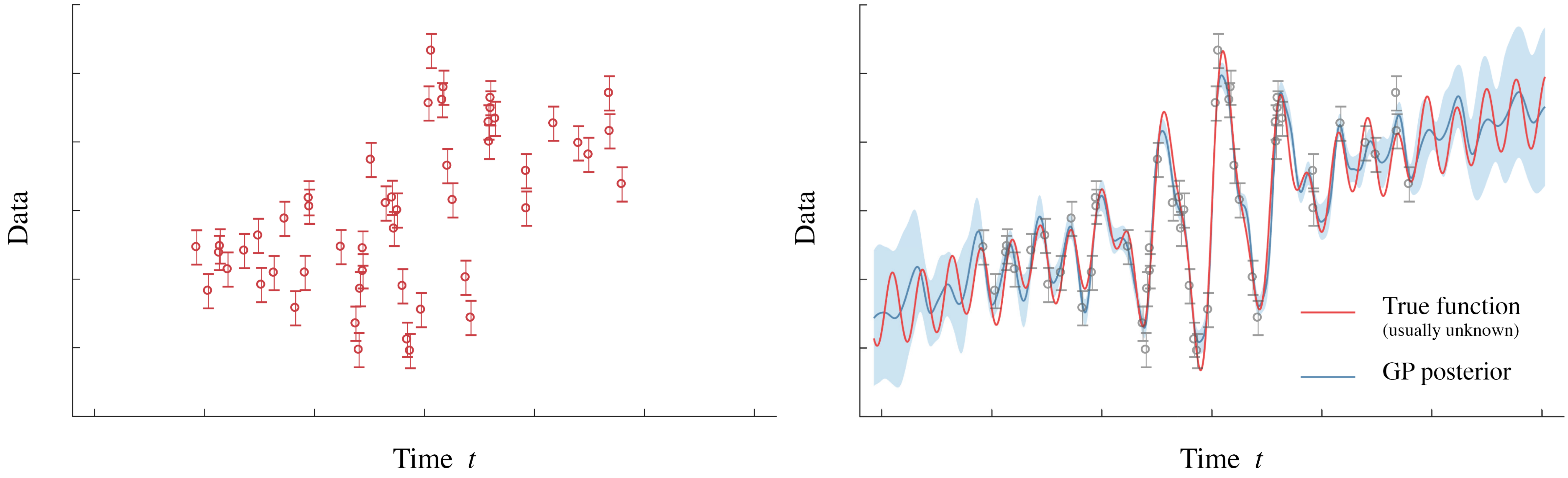
covariance *hyper-parameters*

Then: $\log \mathcal{L}(\phi, \theta) \propto (\boldsymbol{y} - \boldsymbol{\mu})^{\mathsf{T}} \mathbf{K}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}) + \log \det \mathbf{K}$

+ simple linear algebra yields predictive GP distribution (eqs. 2.22-24, Rasmussen & Williams)

INTRODUCING GPs

WHAT DOES THIS LOOK LIKE IN PRACTICE?



INTRODUCING GPs

FUNCTION PROPERTIES VIA COVARIANCES

a. Smoothness/fuzziness → smooth spot growth/decay
b. Input scales, e.g. evolution time/length scales → spot evolution/lifetimes
c. Output scales/amplitudes, e.g. signal & noise variance → spot coverage; shot noise
d. (Quasi)-periodicities → stellar rotation
e. Stationarity/lack thereof (e.g. change points, long-term trends) → long-term cycles
f. Isotropy

INTRODUCING GPs

QUASI-PERIODIC COVARIANCE KERNEL

characteristic period

$$k(\tau) \propto \exp\left(\frac{-\tau^2}{2\lambda_e^2}\right) \exp\left(\frac{-\sin^2(\pi\tau / P)}{2\lambda_p^2}\right),$$

evolution time scale

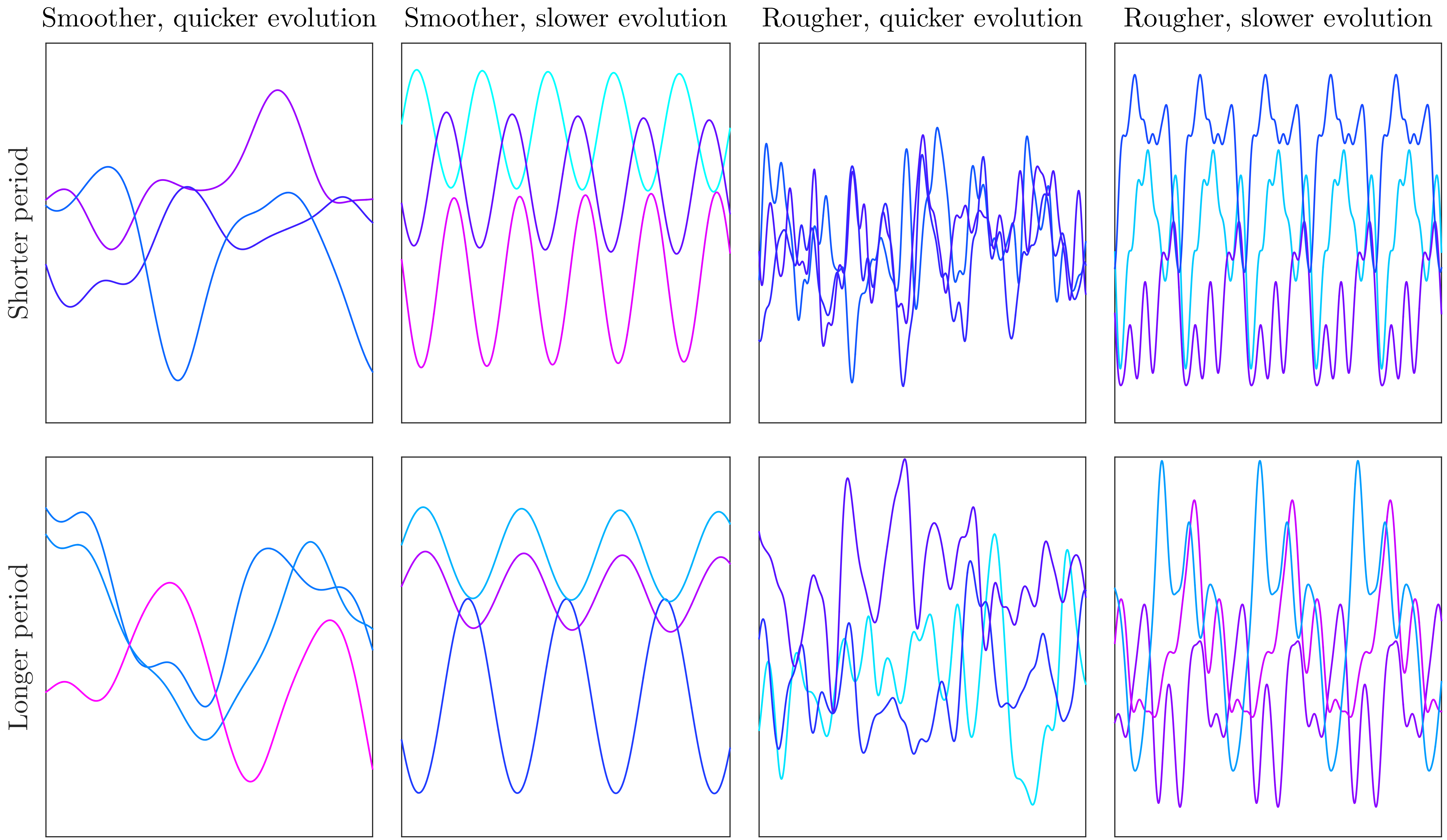
roughness/structure per period

where $\tau := t - t'$

time between any 2 points

INTRODUCING GPs

QUASI-PERIODIC COVARIANCE KERNEL



- Just 3 hyper-parameters yields an enormous diversity of functions
- (Function amplitude would usually be a 4th hyper-parameter)

INTRODUCING GPs

MODELLING STELLAR ACTIVITY *WITHOUT* GPs

[…] fitting sine waves at the rotational period of the star and the significant harmonics […]

The global model fitted on the RVs is therefore:

$$
\begin{aligned}
\text{subset 2008} \quad &: \quad lin0 + lin1 \cdot JDB_{2008} + lin2 \cdot JDB_{2008}^2 + A_{RV-Rhk} \cdot RHK_{low\,freq,2008} \\
\text{subset 2009} \quad &: \quad lin0 + lin1 \cdot JDB_{2009} + lin2 \cdot JDB_{2009}^2 + A_{RV-Rhk} \cdot RHK_{low\,freq,2009} \\
&\quad +A11s \cdot sin(\frac{2\pi}{P1}) \cdot JDB_{2009} + A11c \cdot cos(\frac{2\pi}{P1}) \cdot JDB_{2009} \\
&\quad +A12s \cdot sin(\frac{2\pi}{P1/2}) \cdot JDB_{2009} + A12c \cdot cos(\frac{2\pi}{P1/2}) \cdot JDB_{2009} \\
\text{subset 2010} \quad &: \quad lin0 + lin1 \cdot JDB_{2010} + lin2 \cdot JDB_{2010}^2 + A_{RV-Rhk} \cdot RHK_{low\,freq,2010} \\
&\quad +A21s \cdot sin(\frac{2\pi}{P2}) \cdot JDB_{2010} + A21c \cdot cos(\frac{2\pi}{P2}) \cdot JDB_{2010} \\
&\quad +A23s \cdot sin(\frac{2\pi}{P2/3}) \cdot JDB_{2010} + A23c \cdot cos(\frac{2\pi}{P2/3}) \cdot JDB_{2010} \\
&\quad +A24s \cdot sin(\frac{2\pi}{P2/4}) \cdot JDB_{2010} + A24c \cdot cos(\frac{2\pi}{P2/4}) \cdot JDB_{2010} \\
\text{subset 2011} \quad &: \quad lin0 + lin1 \cdot JDB_{2011} + lin2 \cdot JDB_{2011}^2 + A_{RV-Rhk} \cdot RHK_{low\,freq,2011} \\
&\quad +A31s \cdot sin(\frac{2\pi}{P3}) \cdot JDB_{2011} + A31c \cdot cos(\frac{2\pi}{P3}) \cdot JDB_{2011} \\
&\quad +A32s \cdot sin(\frac{2\pi}{P3/2}) \cdot JDB_{2011} + A32c \cdot cos(\frac{2\pi}{P3/2}) \cdot JDB_{2011} \\
&\quad +A33s \cdot sin(\frac{2\pi}{P3/3}) \cdot JDB_{2011} + A33c \cdot cos(\frac{2\pi}{P3/3}) \cdot JDB_{2011}
\end{aligned}
$$

23 free parameters just for stellar activity

- Parametrising stellar activity signals can be extremely difficult
- Model on left for Alpha Cen B RVs (taken from **Dumusque+12**)

INTRODUCING GPs

QUASI-PERIODIC COVARIANCE KERNEL

characteristic period

$$k(\tau) \propto \exp\left(\frac{-\tau^2}{2\lambda_e^2}\right)\exp\left(\frac{-\sin^2(\pi\tau / P)}{2\lambda_p^2}\right),$$

evolution time scale

roughness/structure per period

INTRODUCING GPs

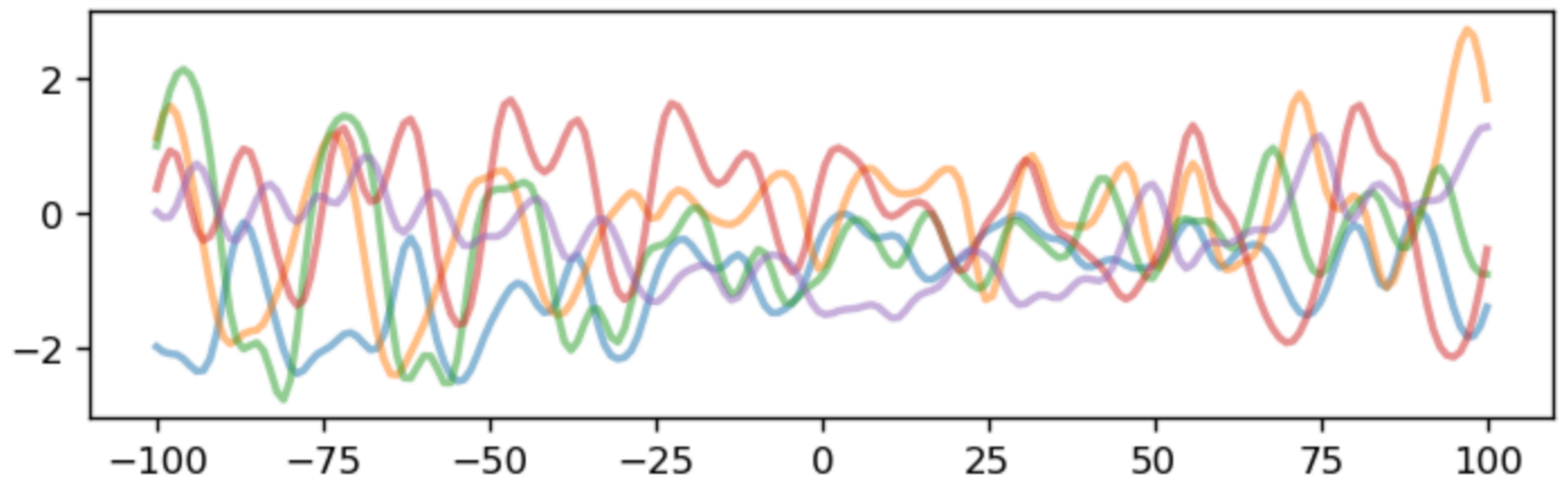
VERY EASY TO IMPLEMENT IN PYTHON

```
[1]: import numpy as np
     import matplotlib.pyplot as plt
     import scipy.linalg as spl
     from numpy.random import multivariate_normal as mvn

     jitter = 1e-10 # small term added to diag(K) for more stable matrix inversion

     def K_QP(t1, t2, theta):
         tau = np.subtract.outer(t1,t2)
         h, P, lambda_p, lambda_e = theta
         K = (h**2)*np.exp(-((np.sin(np.pi*tau/P)**2)/(lambda_p**2) + (tau/lambda_e)**2)/2)
         np.fill_diagonal(K, K.diagonal() + jitter)
         return K
```

```
[2]: plt.figure(num=None, figsize=(7, 2), dpi=120, facecolor='w', edgecolor='k')
     t_obs = np.linspace(-100,100,200)
     for i in range(5):
         plt.plot(t_obs,mvn(0*t_obs, K_QP(t_obs, t_obs, [1, 25, 0.5, 50])), alpha=0.5, lw =2)
```

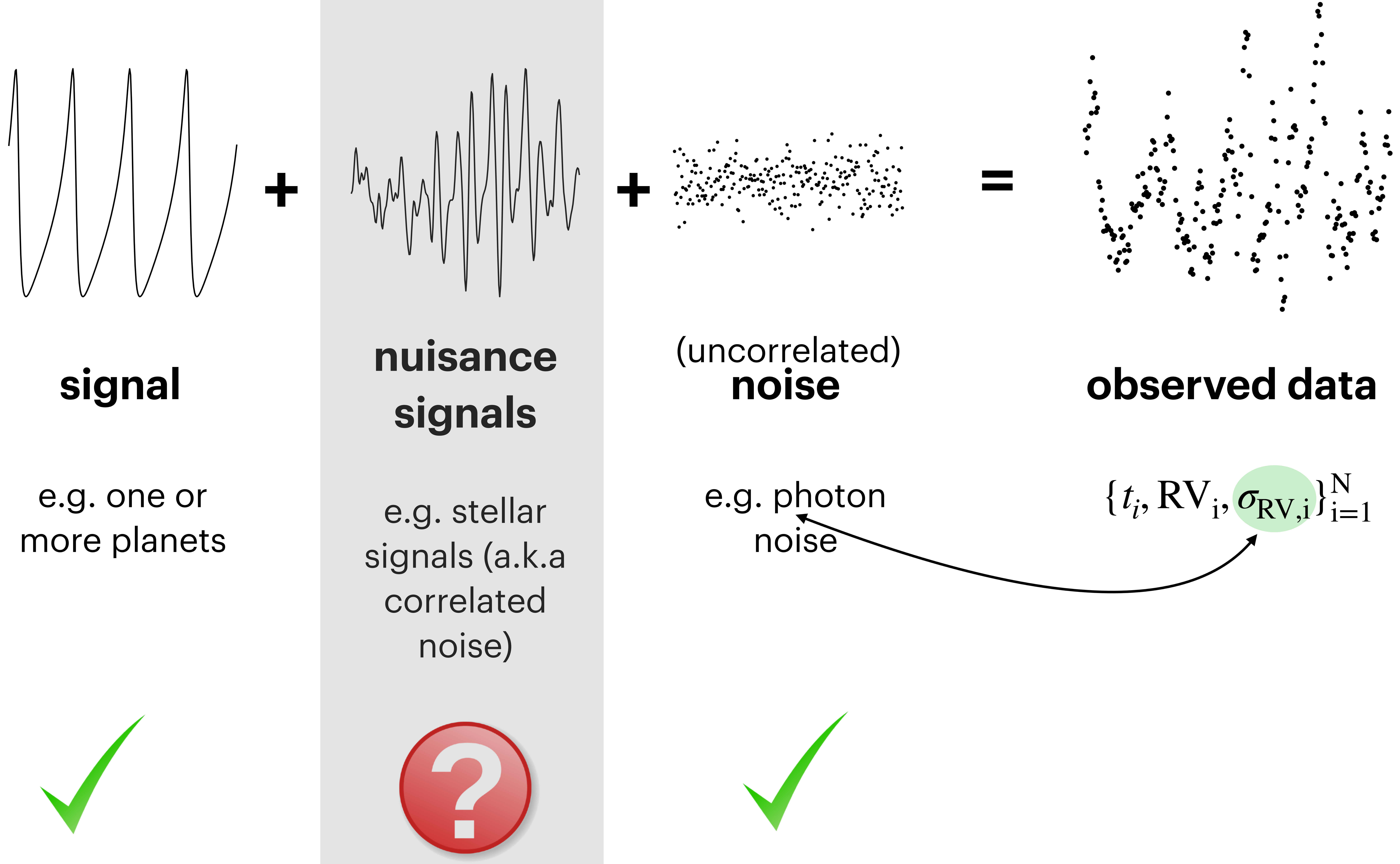



```
[3]: def logL_GP(K, y_res):

         factor, flag = spl.cho_factor(K)
         logdet = 2*np.sum(np.log(np.diag(factor)))
         gof = np.dot(y_res,spl.cho_solve((factor,flag),y_res))
         return -0.5*(gof + logdet + len(y_res)*np.log(2*np.pi))
```

- Writing your own GP code from scratch is easy
- And you'll learn loads
- Code on left sets up quasi-periodic covariance kernel, draws sample functions, computes log likelihood...
- Conditioning on data & prediction also easy
- Python: `GPy`, `sklearn.gaussian_process`, `gpytorch`, `gpflow` ...

APPLICATIONS



+
+
=
signal
nuisance signals
(uncorrelated) noise
observed data
e.g. one or more planets
e.g. stellar signals (a.k.a correlated noise)
e.g. photon noise
$\{t_i, \mathrm{RV}_\mathrm{i}, \sigma_{\mathrm{RV,i}}\}_{\mathrm{i=1}}^{\mathrm{N}}$
?

APPLICATIONS

(I) SEPARATING STELLAR ACTIVITY AND PLANETS

- **Model stellar signals** simultaneously with planet(s) → improved planet detection, characterisation
- Early uses of GPs in this way: CoRoT-7 (**Haywood+14**), Kepler-78 (**Grunblatt+15**)
- Works extremely well when $P_{\mathrm{star}} \not\approx P_{\mathrm{planet}}$

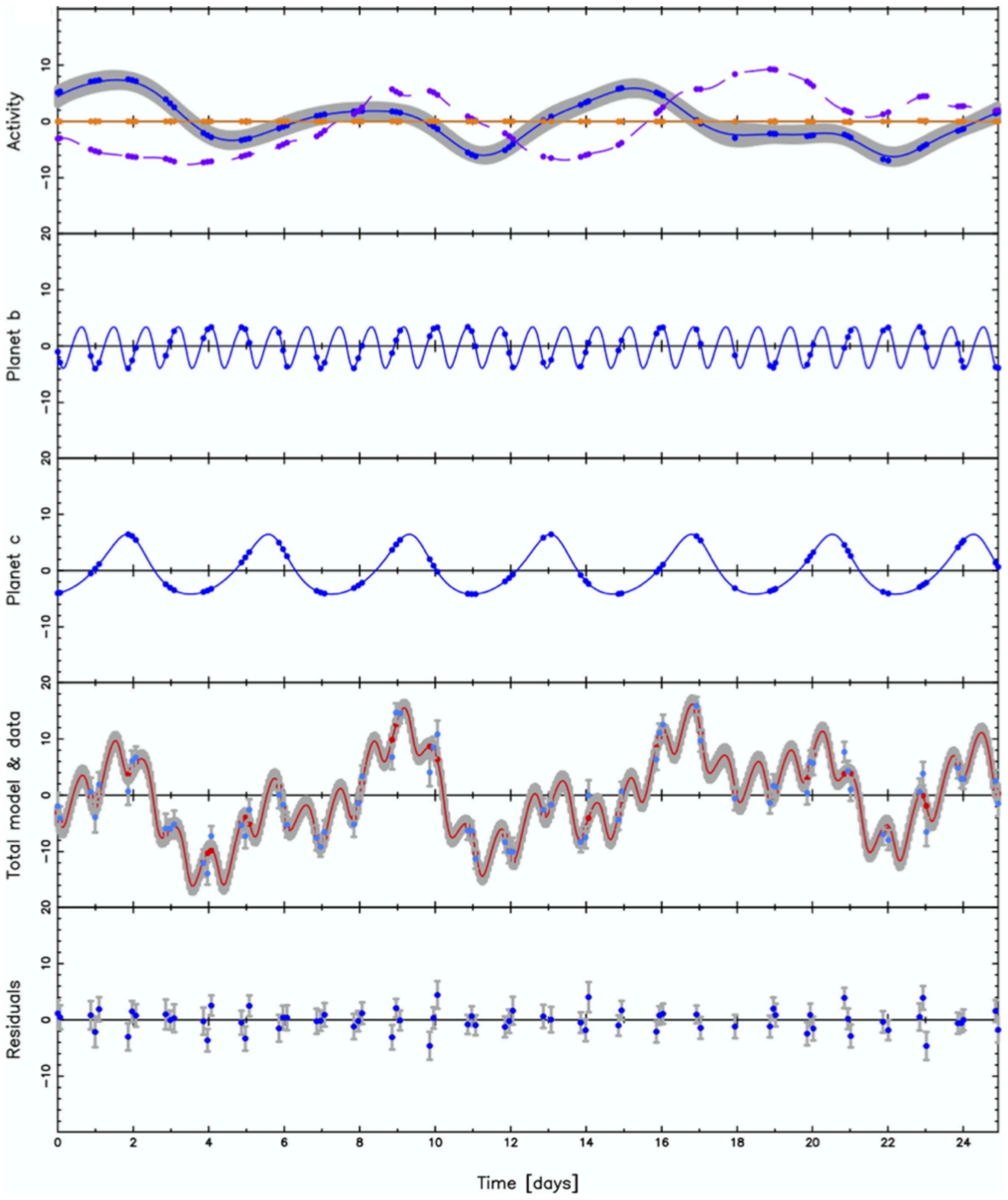
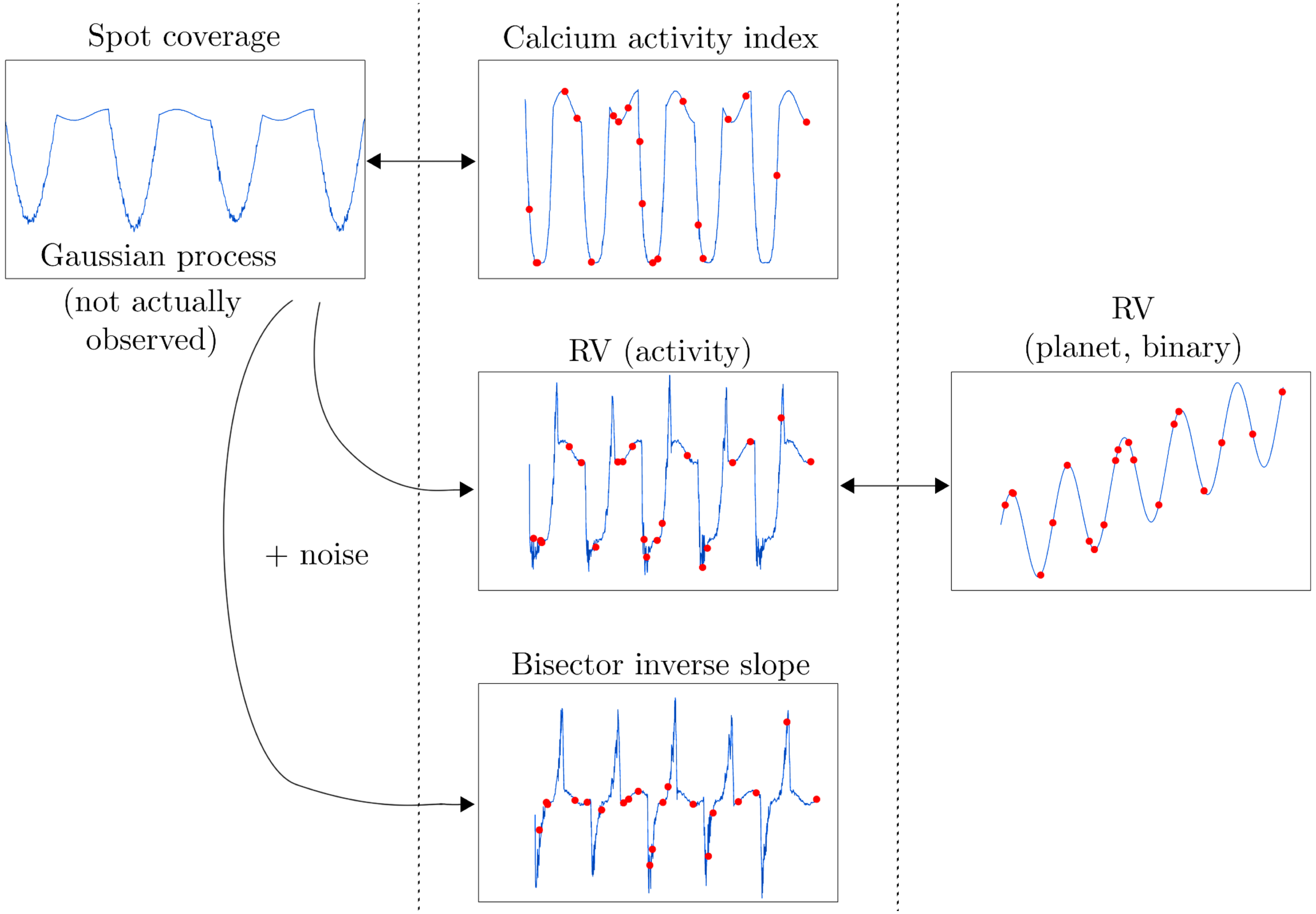


Figure credit **Haywood+14**

APPLICATIONS

(I) SEPARATING STELLAR ACTIVITY AND PLANETS

- Model activity in **RVs + activity diagnostics simultaneously** with a GP (**Rajpaul+15**)
- Improved activity constraints even when $P_{\rm star} \approx P_{\rm planet}$
- Several planets discovered/ characterised in this way



APPLICATIONS

(I) SEPARATING STELLAR ACTIVITY AND PLANETS

- Model activity in **RVs + activity diagnostics simultaneously** with a GP (**Rajpaul+15**)
- Improved activity constraints even when $P_{\text{star}} \approx P_{\text{planet}}$
- Several planets discovered/ characterised in this way

A Gaussian process framework for modelling stellar activity signals in radial velocity data FREE

V. Rajpaul, S. Aigrain, M. A. Osborne, S. Reece, S. Roberts

Monthly Notices of the Royal Astronomical Society, Volume 452, Issue 3, 21 September 2015, Pages 2269–2291, https://doi.org/10.1093/mnras/stv1428

Published: 23 July 2015 **Article history ▾**

An 11 Earth-mass, Long-period Sub-Neptune Orbiting a Sun-like Star

Andrew W. Mayo[1,2,3,22,23], Vinesh M. Rajpaul[4], Lars A. Buchhave[2,3], Courtney D. Dressing[1], Annelies Mortier[4,5], Li Zeng[6,7], Charles D. Fortenbach[8], Suzanne Aigrain[9], Aldo S. Bonomo[10], Andrew Collier Cameron[5] + Show full author list

Published 2019 September 27 • © 2019. The American Astronomical Society. All rights reserved.

The Astronomical Journal, Volume 158, Number 4

APPLICATIONS

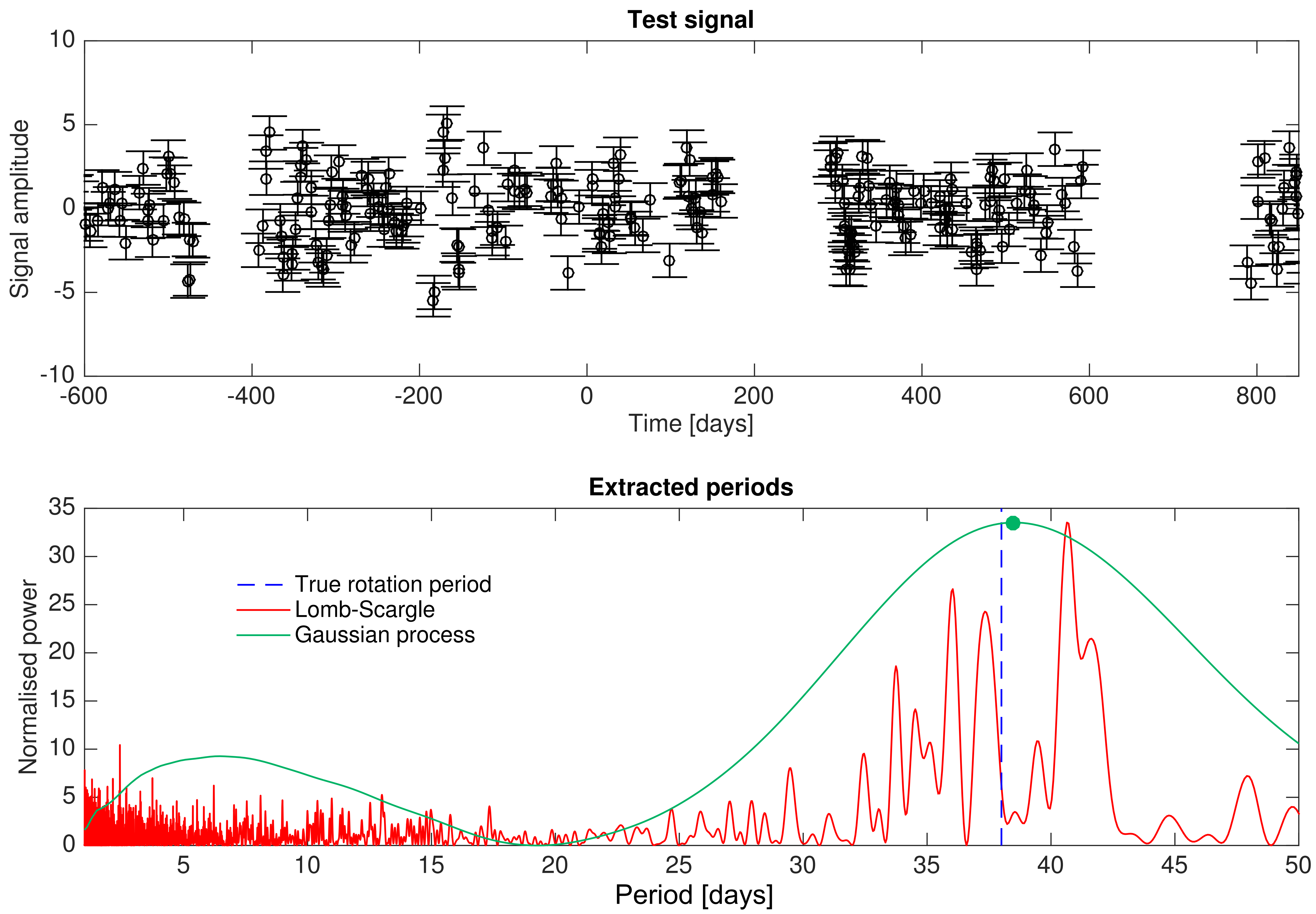
(II) STUDYING PERIODIC PHENOMENA

- Lomb-Scargle periodogram: OK for single sinusoids + white noise
- What about multiple non-sinusoidal, quasi-periodic signals + correlated noise?
- GP can provide drop-in replacement!

$$k(\tau) \propto \exp\left(\frac{-\tau^2}{2\lambda_e^2}\right)\exp\left(\frac{-\sin^2(\pi\tau / P)}{2\lambda_p^2}\right)$$

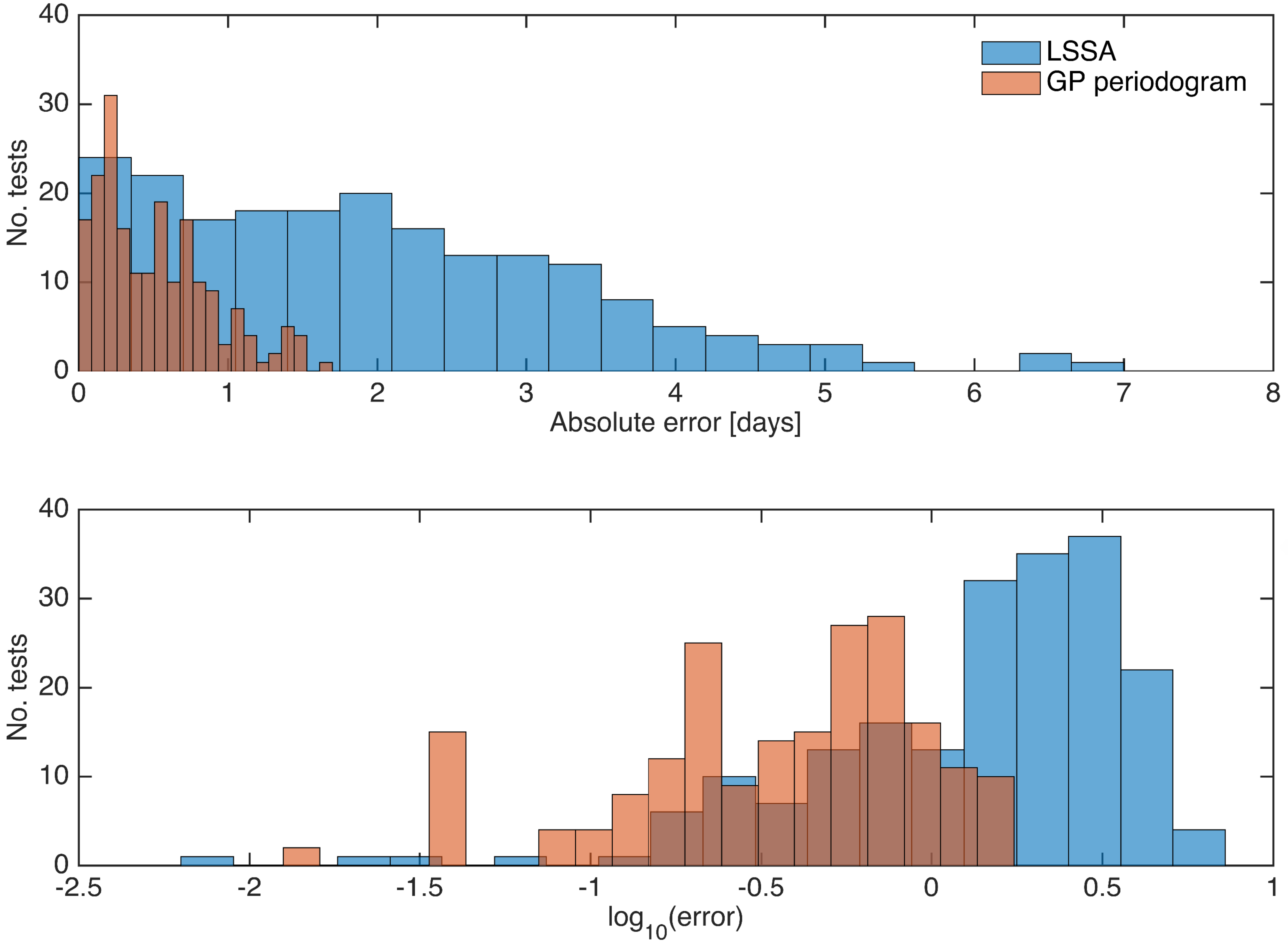
APPLICATIONS

(II) STUDYING PERIODIC PHENOMENA



APPLICATIONS

(II) STUDYING PERIODIC PHENOMENA



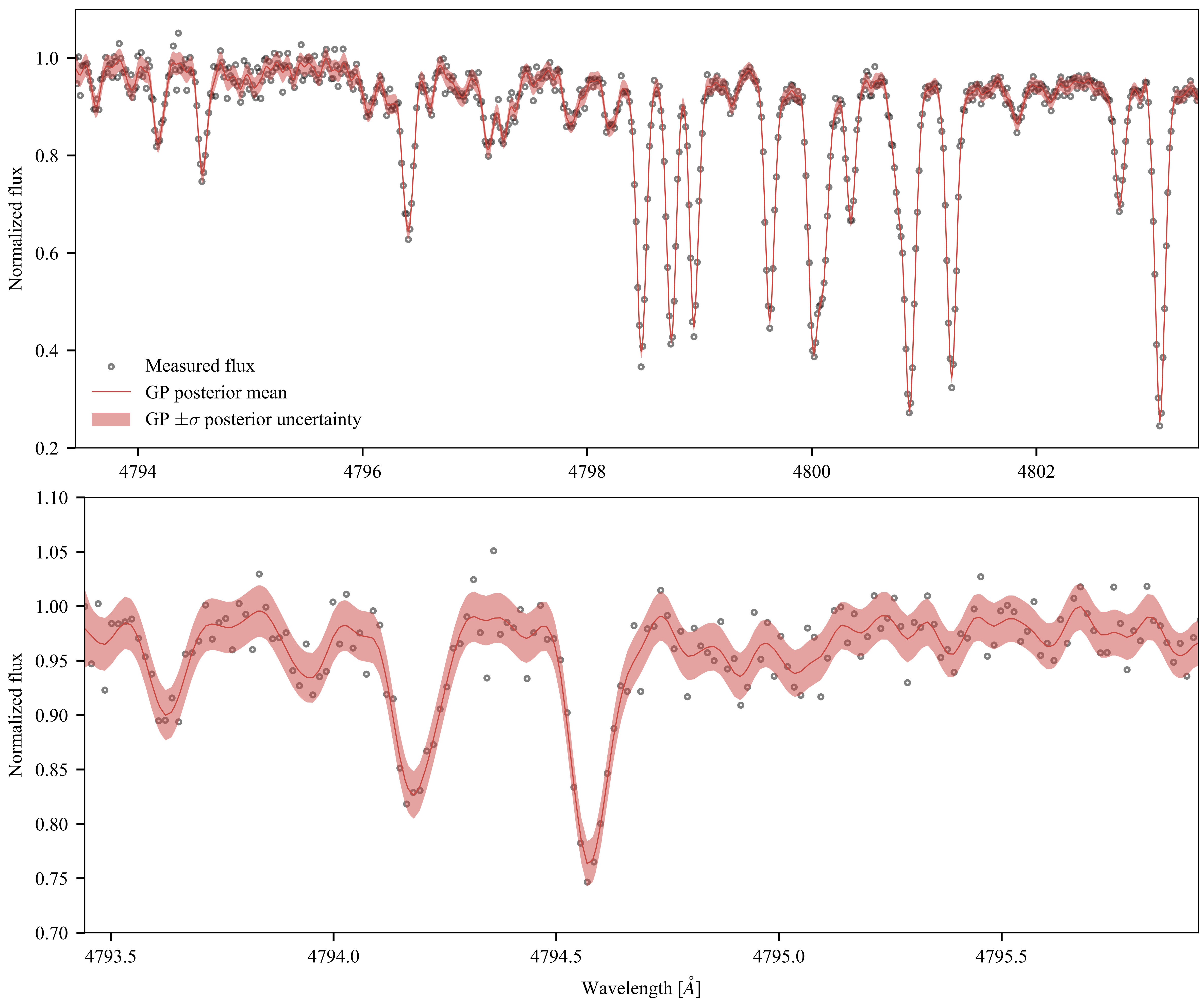
Inferring probabilistic stellar rotation periods using Gaussian processes FREE

Ruth Angus, Timothy Morton, Suzanne Aigrain, Daniel Foreman-Mackey, Vinesh Rajpaul

Monthly Notices of the Royal Astronomical Society, Volume 474, Issue 2, February 2018, Pages 2094–2108, https://doi.org/10.1093/mnras/stx2109

Published: 22 September 2017 **Article history** ▾

APPLICATIONS



(III) MODELLING SPECTRA FOR RV EXTRACTION

- Modelling stellar spectra: usually a lot of work & astrophysical input
- GPs → very good, non-parametric spectra; almost zero effort
- Leads to simple RV extraction
- Can also reduce stellar activity contamination (e.g. Kepler-37 - with Buchhave, Aigrain+, *in prep.*)

APPLICATIONS

Disentangling Time-series Spectra with Gaussian Processes: Applications to Radial Velocity Analysis

Ian Czekala[1,7], Kaisey S. Mandel[2], Sean M. Andrews[2], Jason A. Dittmann[2], Sujit K. Ghosh[3,4], Benjamin T. Montet[5,8], and Elisabeth R. Newton[6,9]

Published 2017 May 4 •

The Astrophysical Journal, Volume 840, Number 1

A robust, template-free approach to precise radial velocity extraction

V M Rajpaul, S Aigrain, L A Buchhave

Monthly Notices of the Royal Astronomical Society, Volume 492, Issue 3, March 2020, Pages 3960–3983, https://doi.org/10.1093/mnras/stz3599

Published: 03 January 2020 **Article history** ▾

(III) MODELLING SPECTRA FOR RV EXTRACTION

- Modelling stellar spectra: usually a lot of work & astrophysical input
- GPs → very good, non-parametric spectra; almost zero effort
- Leads to simple RV extraction
- Can also reduce stellar activity contamination (e.g. Kepler-37 - with Buchhave, Aigrain+, *in prep.*)

APPLICATIONS

(IV) MODELLING INSTRUMENTAL SYSTEMATICS

- Instruments are not always well-behaved!
- Case-in-point: Kepler's K2 mission
- Right: K2SC's GP modelling of instrumental, astrophysical variability

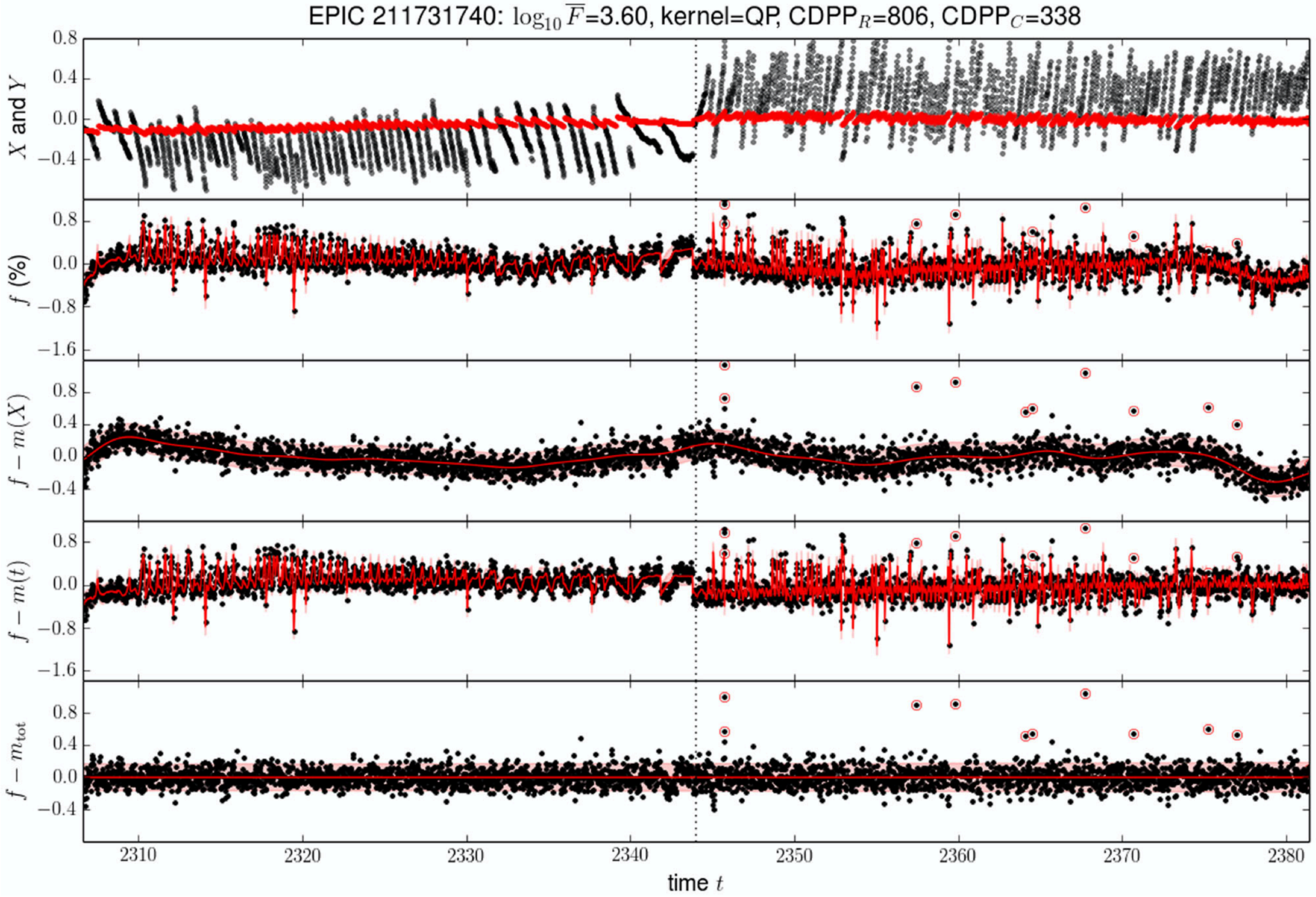


Figure credit **Aigrain+16**

APPLICATIONS

(V) A POWERFUL SIMULATION TOOL

- Train GP on **real data** (in all its messy glory)
- Generate **realistic synthetic data** (same covariance properties; arbitrary sampling)
- Use e.g. to study **observing strategies** and **detection limits**
- Or to identify **artefacts** associated with **fitted models** and **discrete sampling**

APPLICATIONS

(V) A POWERFUL SIMULATION TOOL

- Notable application: showing Alpha Cen Bb was a false positive

Ghost in the time series: no planet for Alpha Cen B

FREE

V. Rajpaul, S. Aigrain, S. Roberts

Monthly Notices of the Royal Astronomical Society: Letters, Volume 456, Issue 1, 11 February 2016, Pages L6–L10, https://doi.org/10.1093/mnrasl/slv164

Published: 20 November 2015 **Article history** ▾

limitations of

~~DANGER~~

GAUSSIAN PROCESSES

LIMITATIONS

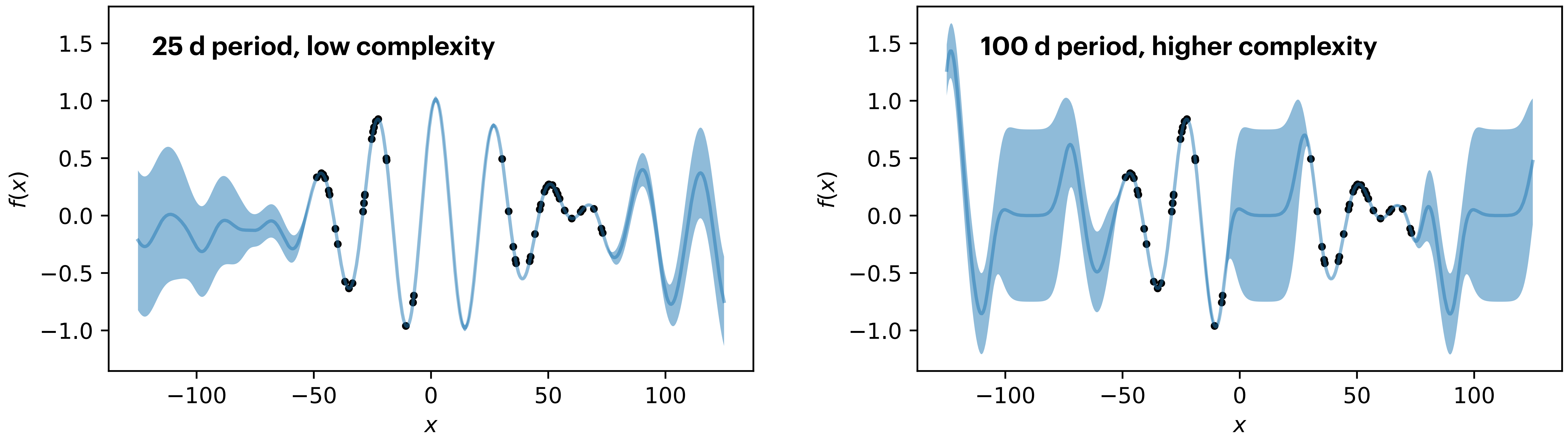
(I) COMPUTATIONALLY EXPENSIVE

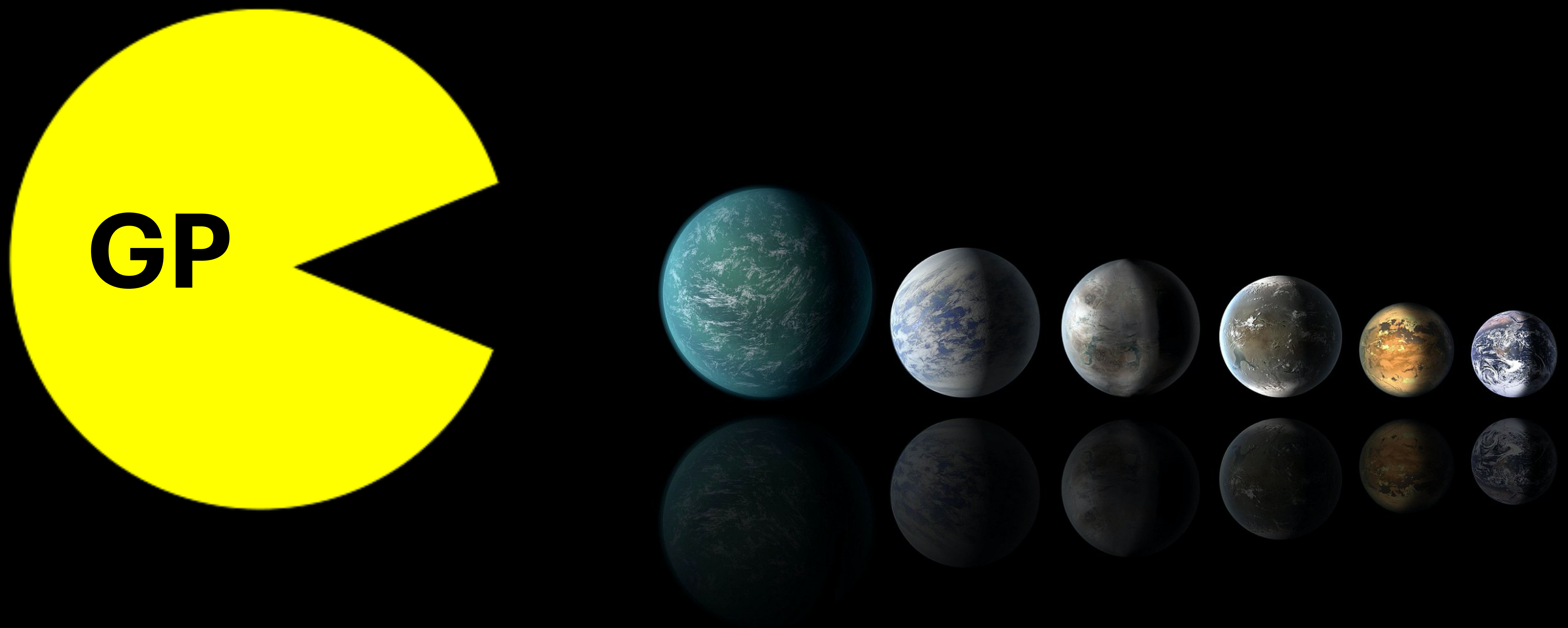
- Must compute $\mathbf{K}^{-1}$ where $\mathbf{K} \in \mathbb{R}^{N \times N}$: scales as $\mathcal{O}(N^3)$
- Things get really difficult when $N \gtrsim 1000$
- Clever techniques can often compute stuff much faster: e.g. `george` and `celerite` by DFM *et al.*; see also `gpflow`, `gpytorch`
- Active research on fast techniques, e.g. using sparse approximations

LIMITATIONS

(II) SENSIBLE COVARIANCE KERNEL ESSENTIAL

- Choosing **a sensible kernel is essential**
- Hyper-parameters need **reasonable priors** & **careful interpretation**.
  E.g. for QP kernel, what does P even mean when $\lambda_e < P$?
- Posteriors may be **multimodal** and **degenerate** - thorough sampling needed




GP

LIMITATIONS

(III) ENSURING YOU FIT WHAT YOU MEAN TO FIT

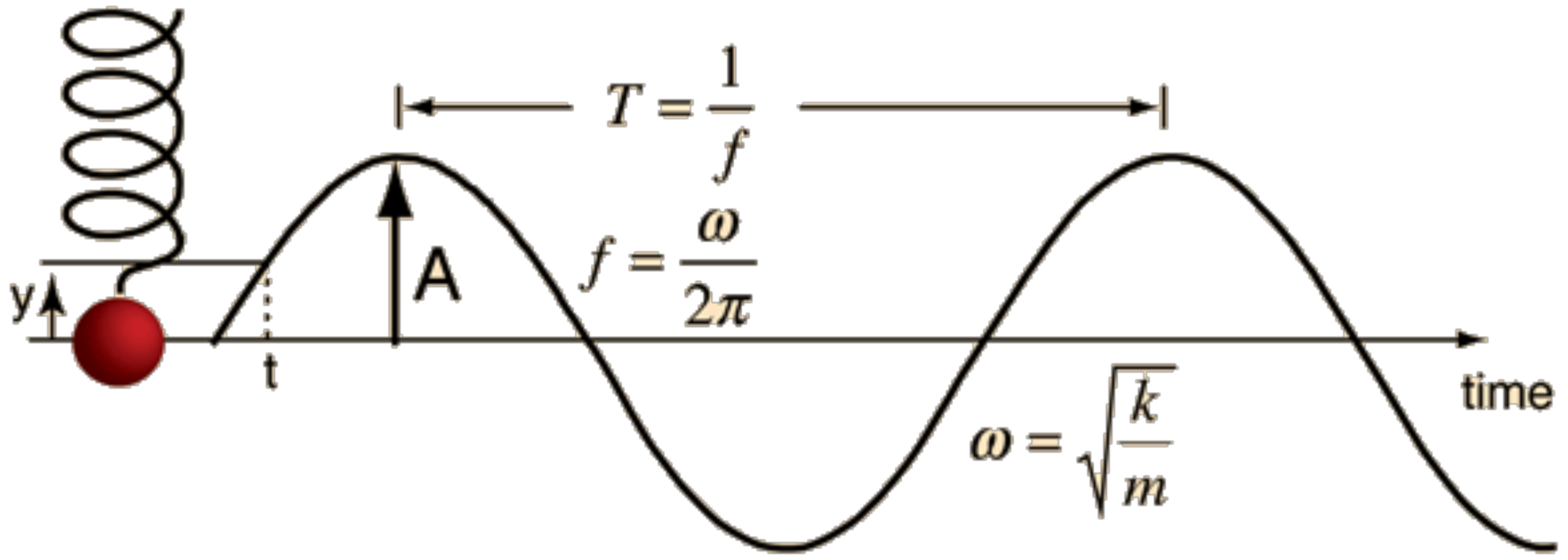
- Can GPs "absorb" planetary signals? Yes...*if* used rashly
- Need good priors + **Bayesian model comparison**
- Use ancillary information + RVs to constrain GP
- Conversely, can stellar activity wrongly be modelled as planets? Yes! Using a GP can avoid such problems. (See **Ahrer+ poster**, forthcoming paper)

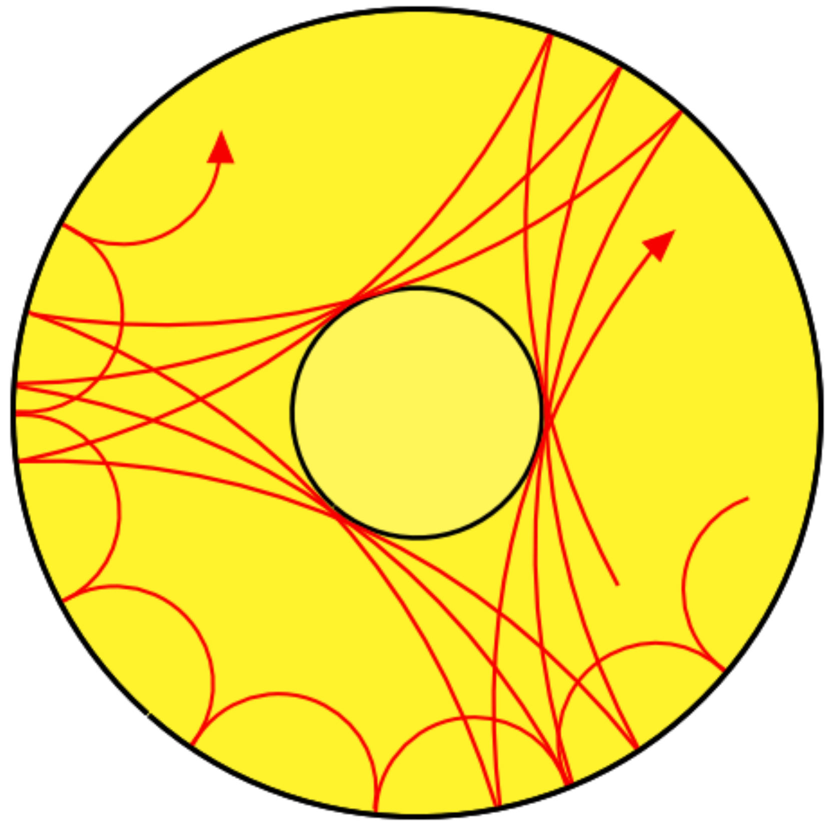

right tool
for the job?
bad workman
blaming his
tools?

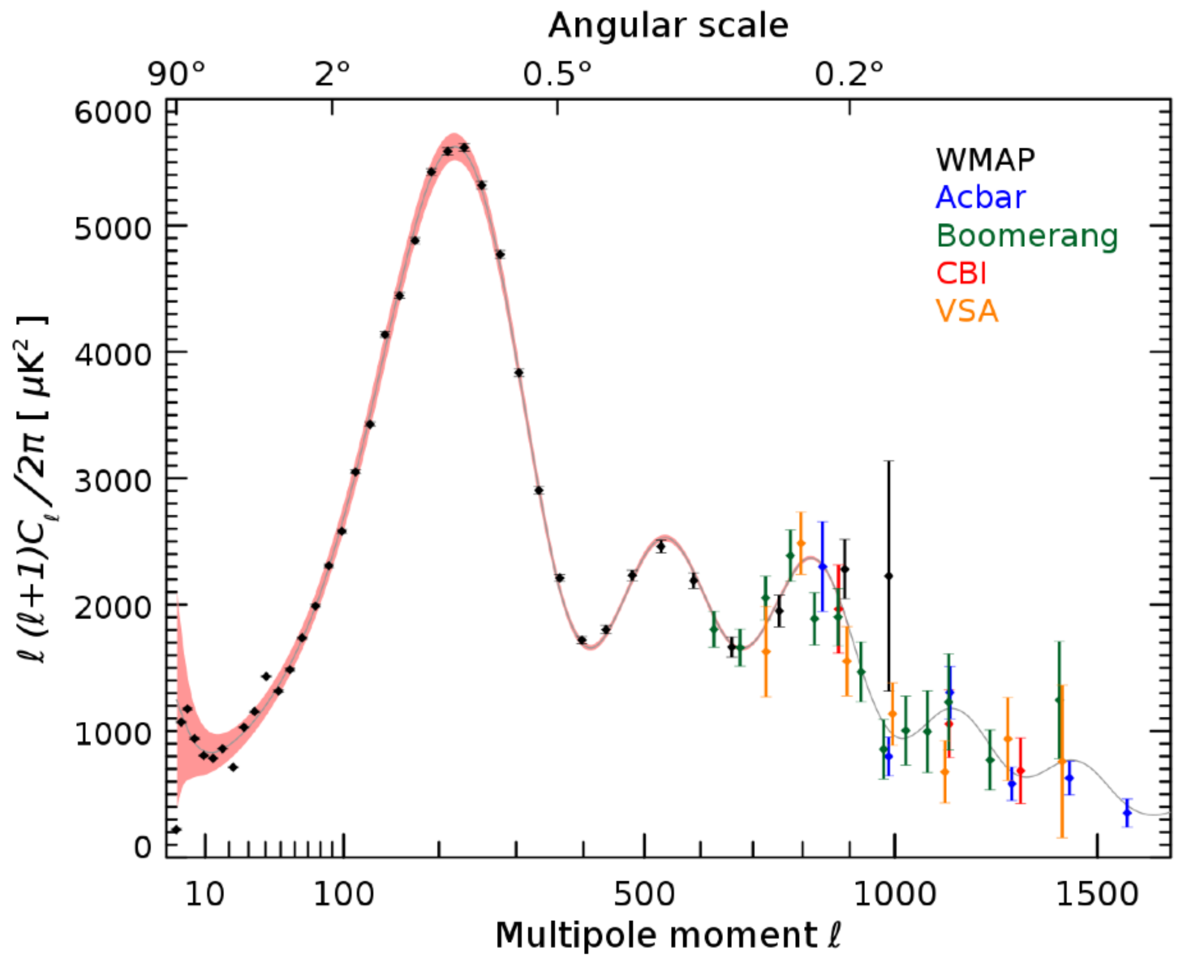
LIMITATIONS

(IV) NOT ALWAYS THE RIGHT TOOL FOR THE JOB

- If a reasonably good **parametric model** exists, use that instead!







- If **Gaussianity is violated** (e.g. due to outliers, heavy-tailed noise process), might need to transform data...or use e.g. a **Student-*t* process**

Q = mcΔt
E = mc²
a² - b²
cos α + cos β = 2
sin α + sin β = 2sin
f(x) = a (x - x₁)
log꜀ a

SUMMARY

HOW TO CHARACTERISE A GP?

- **∞-dimensional** version of a **Gaussian**
- Powerful way to formulate **prior distributions over functions**
- Flexible **Bayesian inference about functions**
  (learn unknown functions + error bars, given data + prior assumptions)

SUMMARY

WHY ARE GPs GREAT?

- **Data-driven Bayesian inference** about **functions**
- Extremely **flexible and powerful**
- Priors can (usually) be **informed by physics**
- Fully **probabilistic**: they "know what they don't know"
- **Easy** to implement
- Often a lot better than the (practical) alternatives

SUMMARY

WHAT ARE THEIR LIMITATIONS?

- Can be **computationally expensive** - usually $\mathcal{O}(N^3)$
- Choice of **covariance function is critical;** hyper-parameters can be **degenerate** and/or **nontrivial to interpret**
- They **can fit anything** - *if* you're not careful
- They're often but **not always the right tool** for the job!

FURTHER READING

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

rsta.royalsocietypublishing.org

Research

CrossMark
click for updates

Cite this article: Roberts S, Osborne M, Ebden M, Reece S, Gibson N, Aigrain S. 2013 Gaussian processes for time-series modelling. Phil Trans R Soc A 371: 20110550.
http://dx.doi.org/10.1098/rsta.2011.0550

Gaussian processes for time-series modelling

S. Roberts[1], M. Osborne[1], M. Ebden[1], S. Reece[1], N. Gibson[2] and S. Aigrain[2]

[1]Department of Engineering Science, and
[2]Department of Astrophysics, University of Oxford, Oxford OX1 3PU, UK

In this paper, we offer a gentle introduction to Gaussian processes for time-series data analysis. The conceptual framework of Bayesian modelling for time-series data is discussed and the foundations of Bayesian non-parametric modelling presented for *Gaussian processes*. We discuss how domain knowledge influences design of the Gaussian process models and provide case examples to highlight the approaches.

Gaussian process tools for modelling stellar signals and studying exoplanets

Vinesh Maguire Rajpaul
Merton College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Trinity 2017

The end