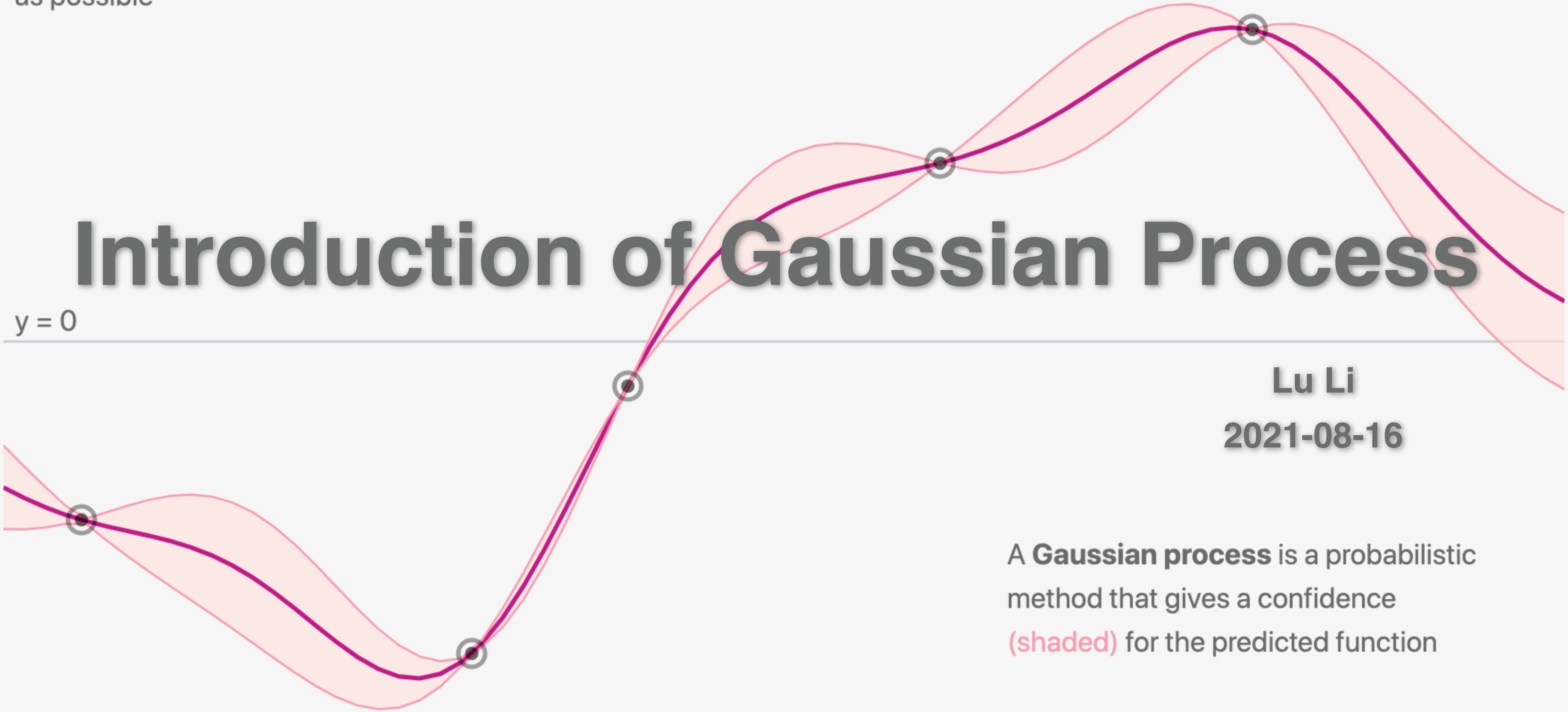
Regression is used to find a function (line) that represents a set of data points as closely as possible

Introduction of Gaussian Process



Lu Li

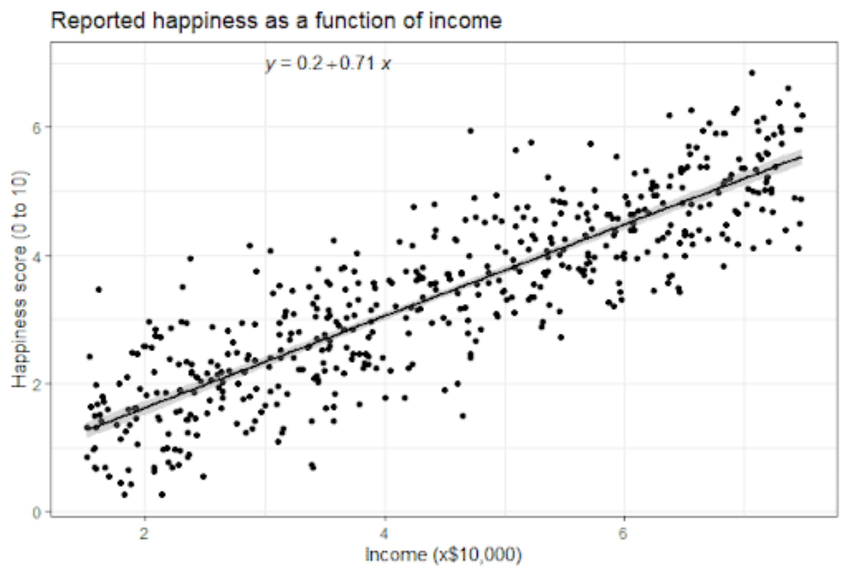
2021-08-16

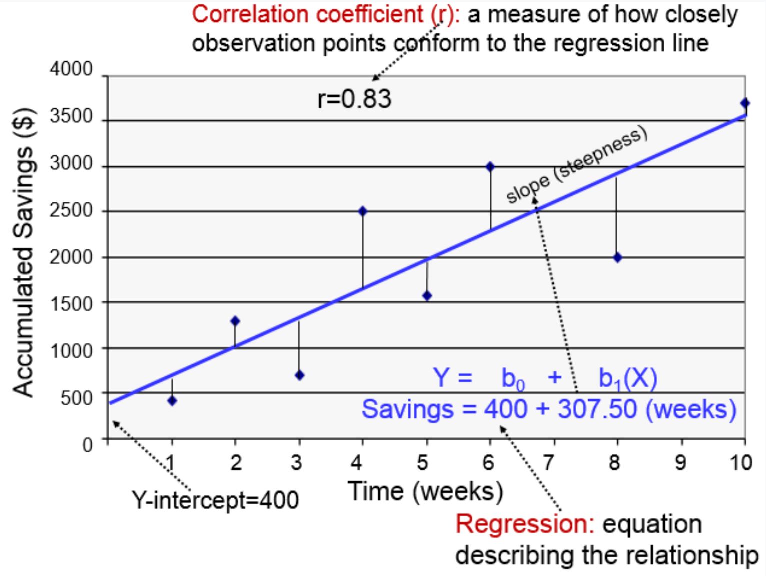
A **Gaussian process** is a probabilistic method that gives a confidence (shaded) for the predicted function

Regression

Definition (wiki)

Regression analysis is a set of statistical processes for estimating the relationships between a dependent variable (often called the 'outcome' or 'response' variable. 因变量，y.) and one or more independent variables (often called 'predictors', 'covariates', 'explanatory variables' or ‘features’. 自变量，x).
Then you can do prediction: a y for a new x.



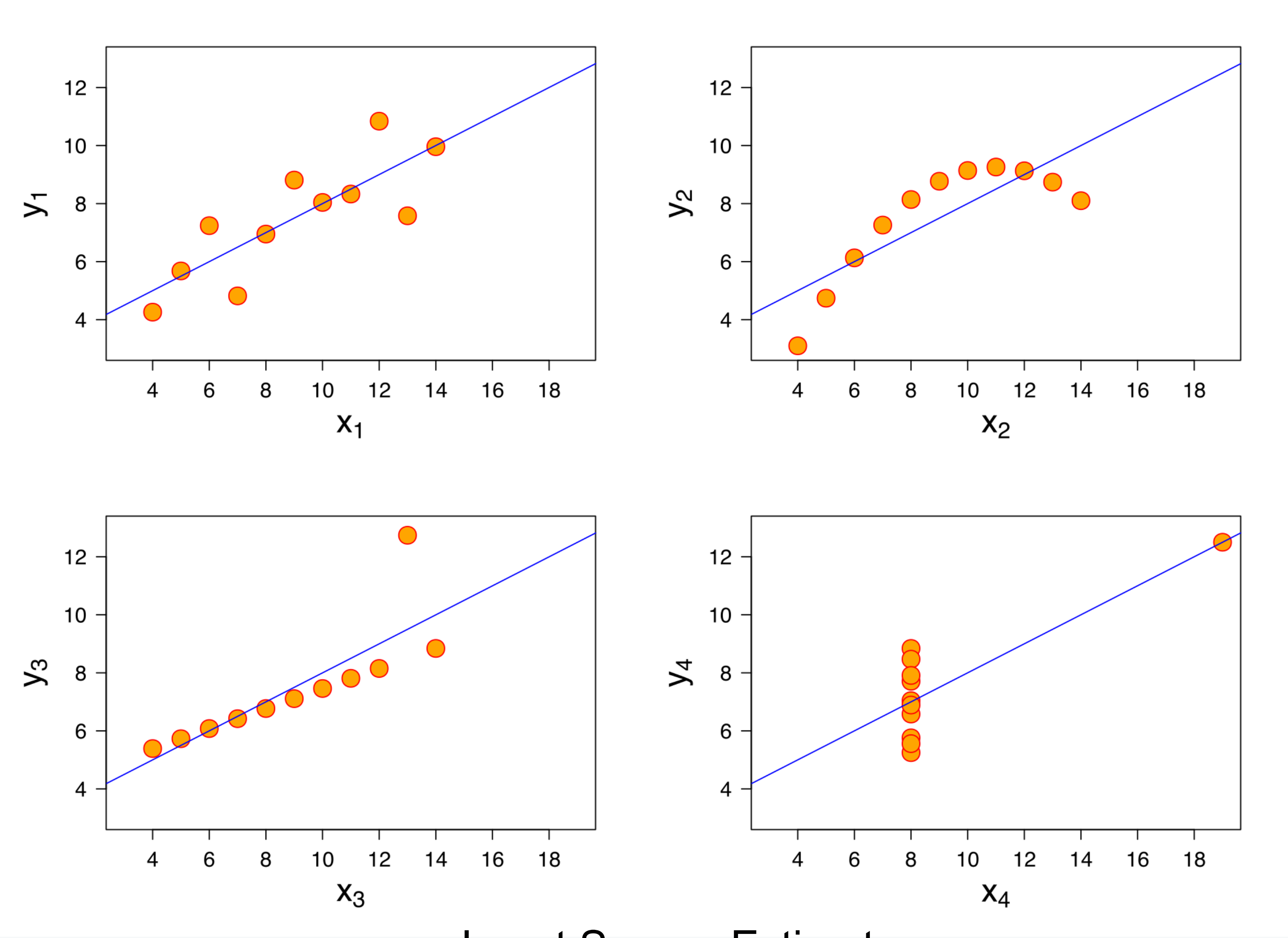


Simple Linear Regression

the same linear regression line, but are graphically very different.

This illustrates the pitfalls of relying solely on a fitted model to understand the relationship between variables.

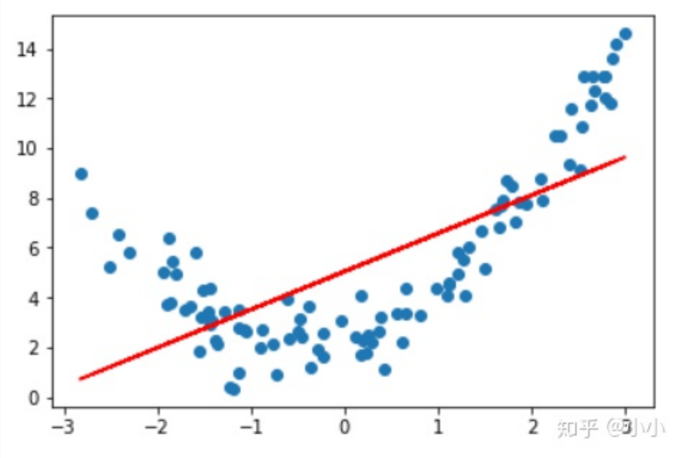
不同的数据得到同样的结果：线性回归，模型依赖的缺陷。

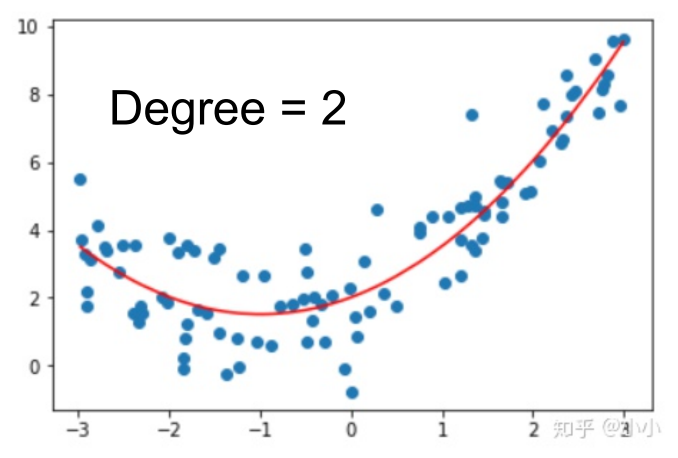


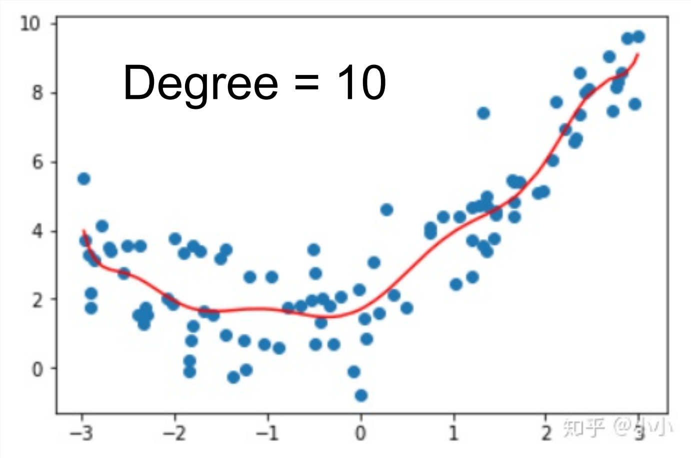
Least Square Estimates

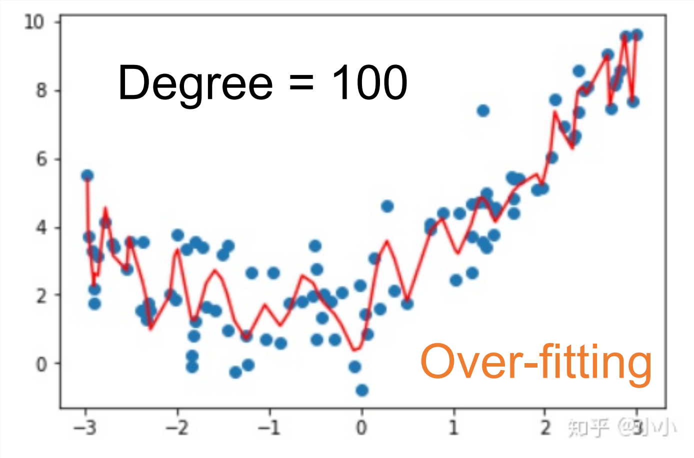
Wikipedia

Polynomial Regression 多项式回归







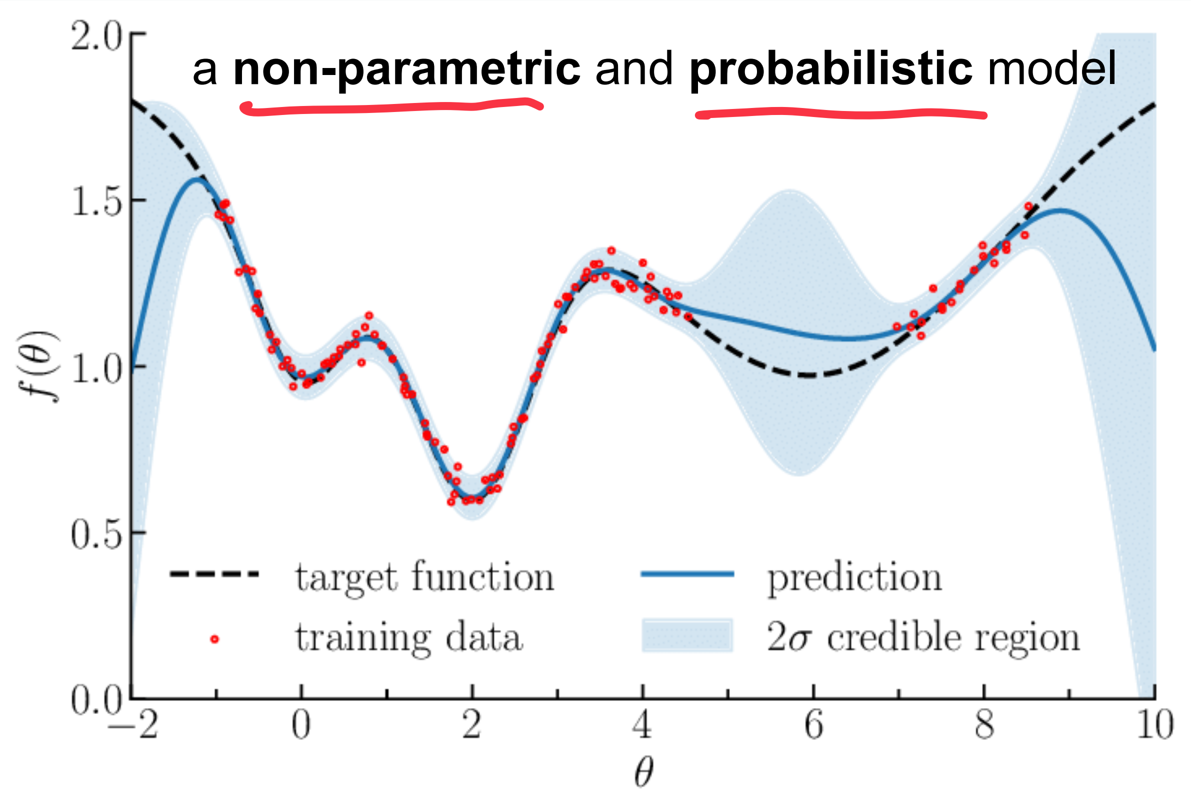


Polynomial regression is a form of regression analysis in which the relationship between the independent variable *x* and the dependent variable *y* is modelled as an *n*th degree polynomial in *x*.

多项式回归是回归分析的一种形式，其中自变量 x 和因变量 y 之间的关系被建模为关于 x 的 n 次多项式

- Parametric, model dependent.
- Can not obtain the confidence level of the “line”.

Gaussian Process (GP)



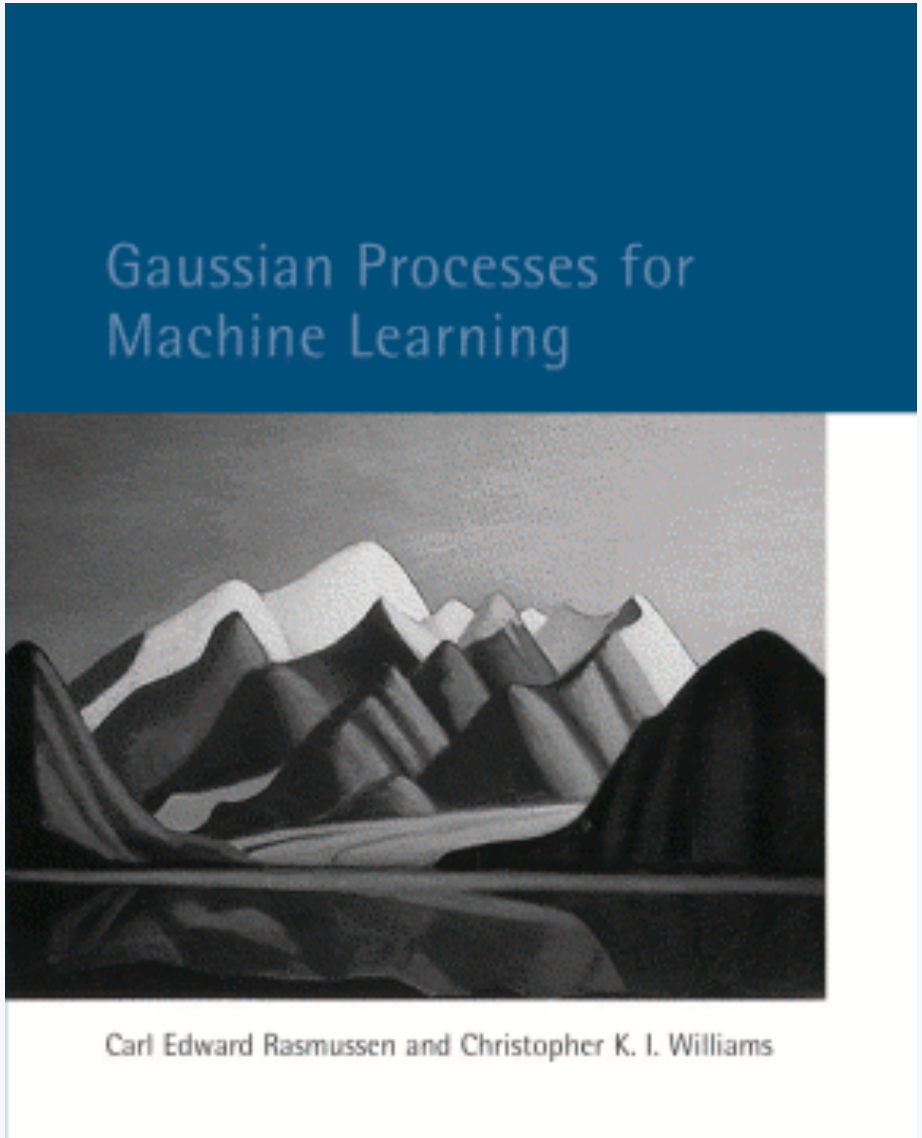
The posterior mean and the uncertainty.

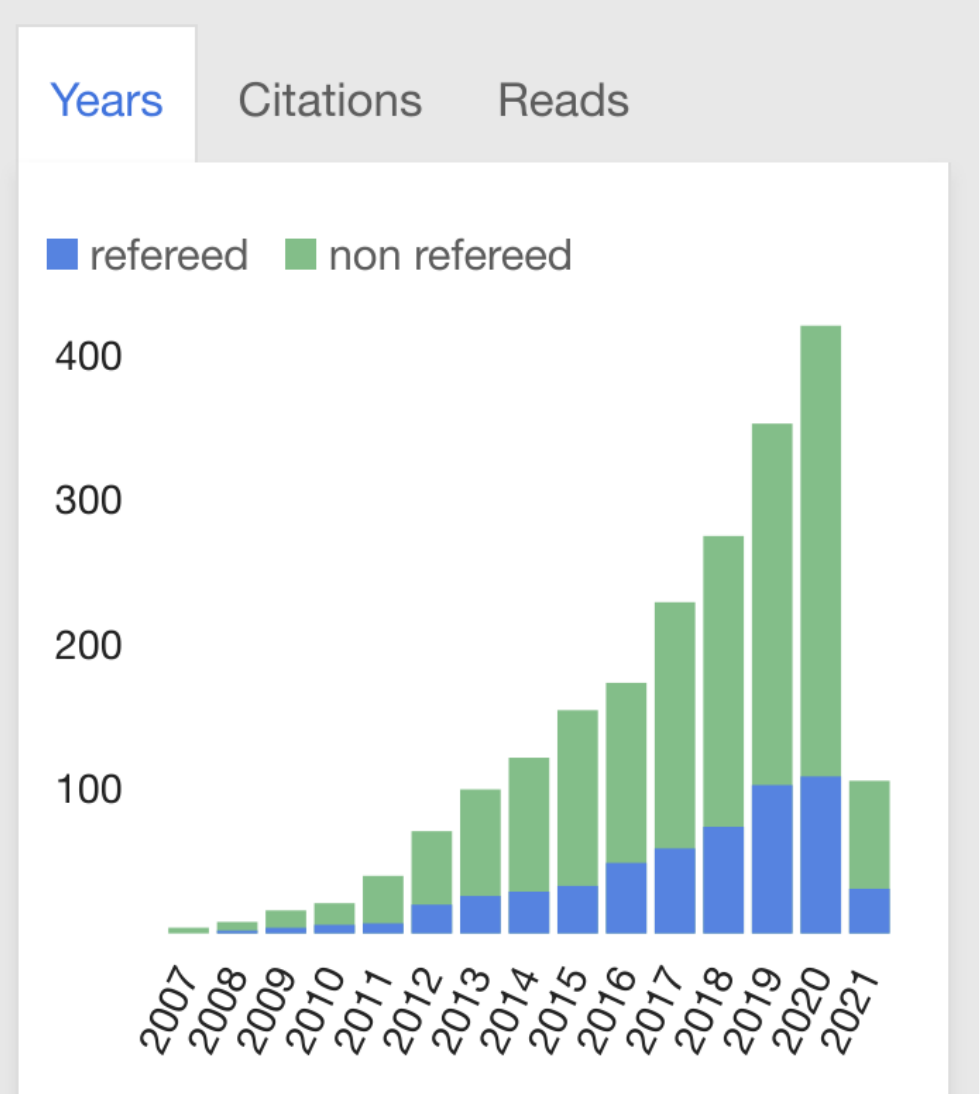
Definition (wiki)

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.

因变量 y_i 服从 multivariate normal distribution

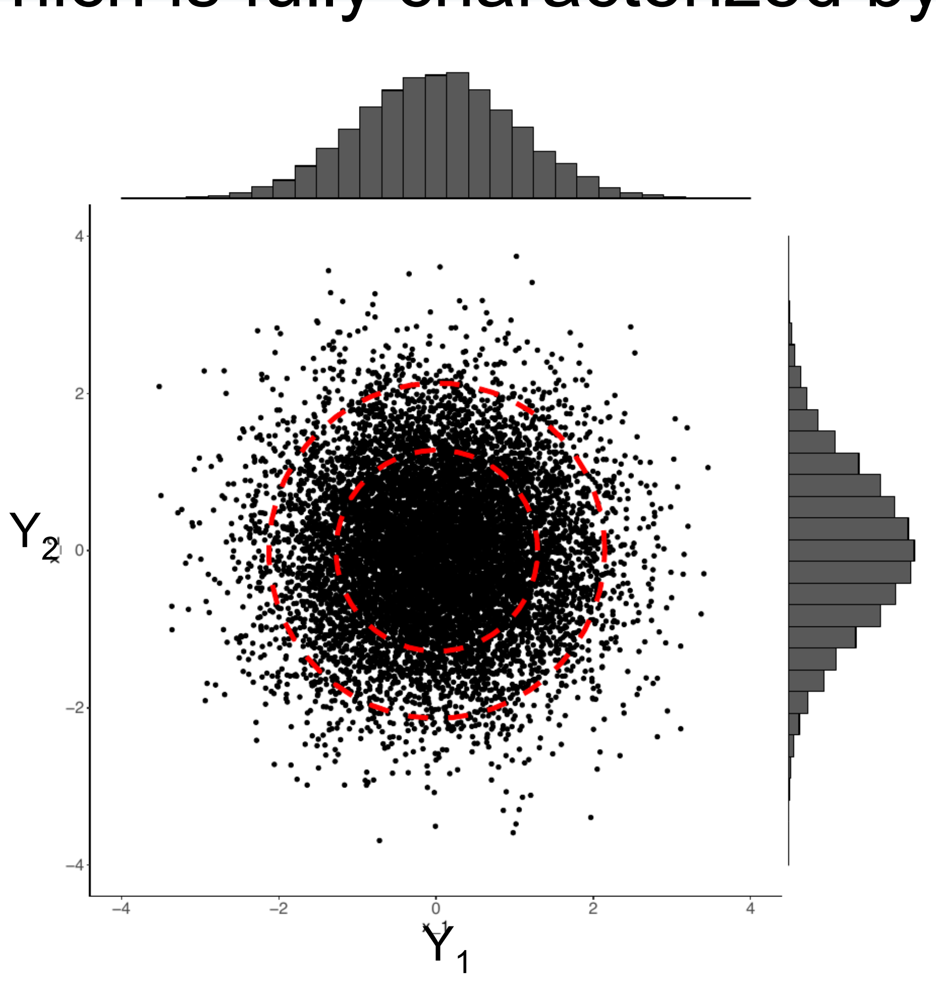
- Data-driven
- Bayesian frame

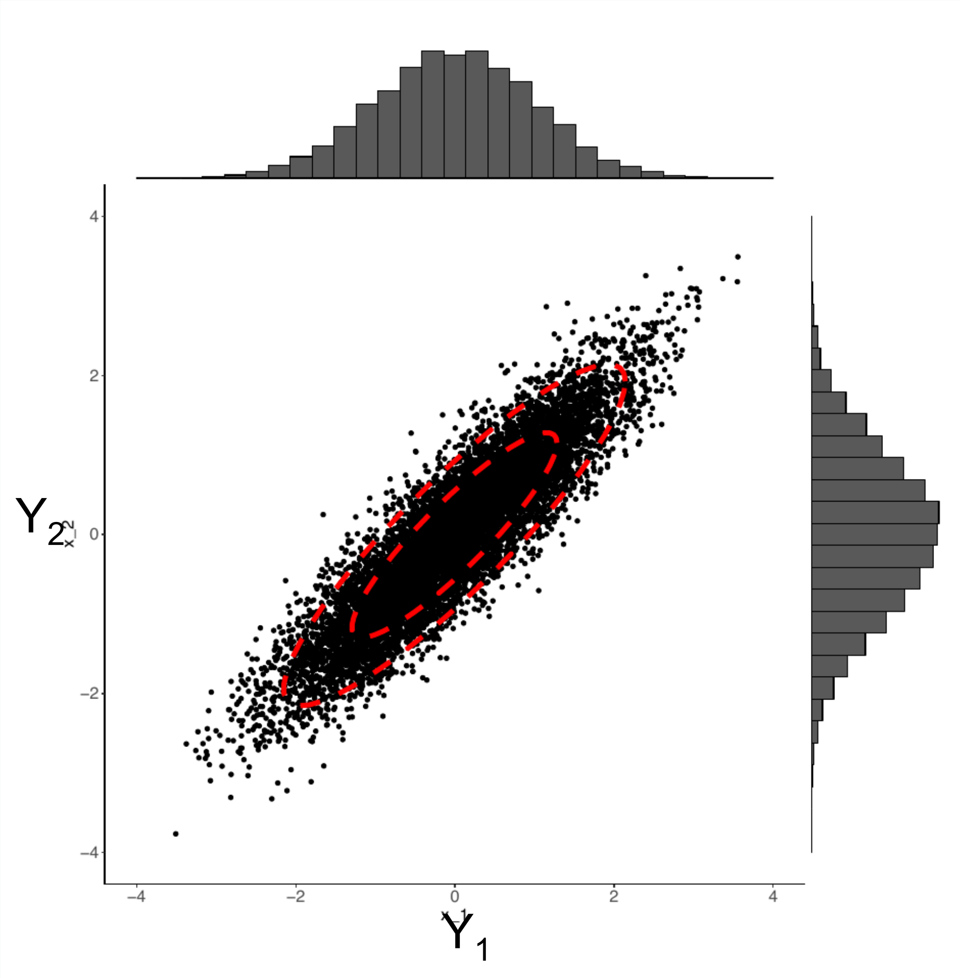

Gaussian Processes for
Machine Learning
Carl Edward Rasmussen and Christopher K. I. Williams


Years
Citations
Reads
refereed
non refereed
400
300
200
100
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021

Multivariate Gaussian distribution

which is fully characterized by mean and covariance matrix



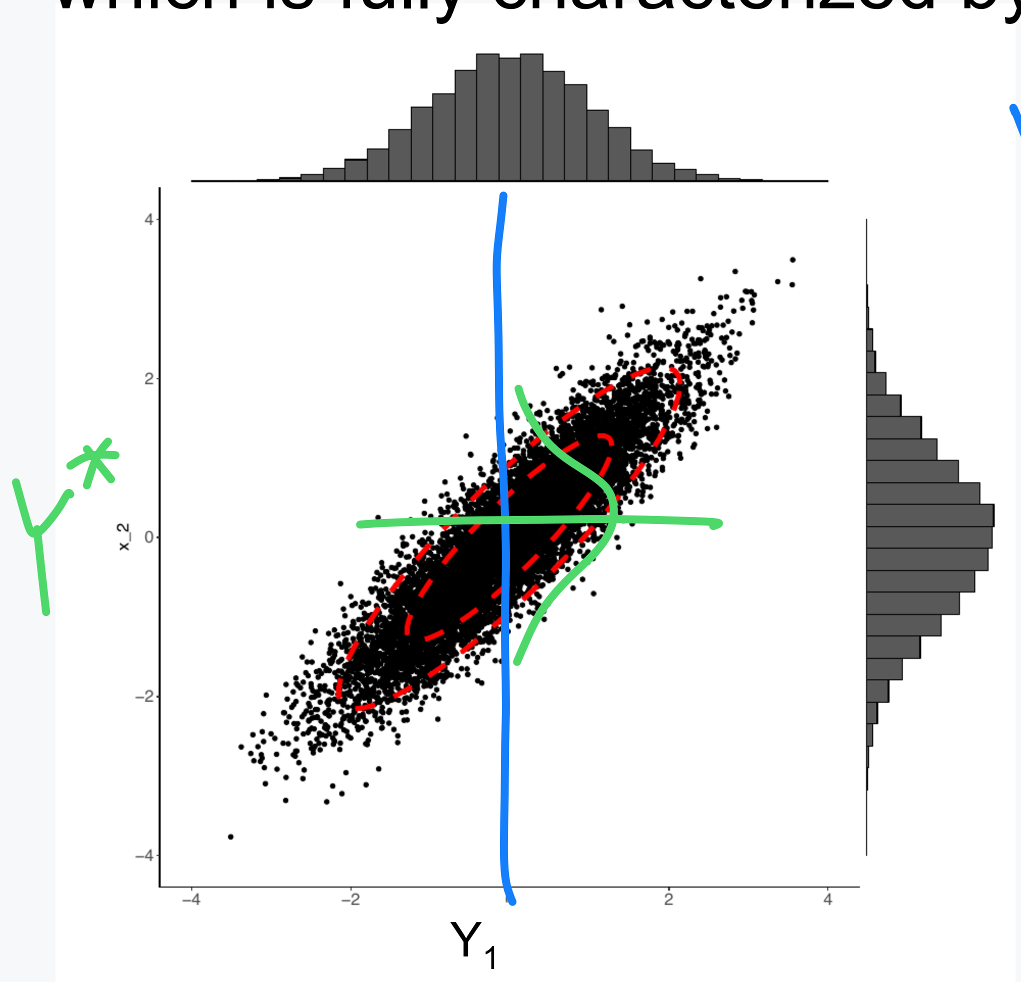


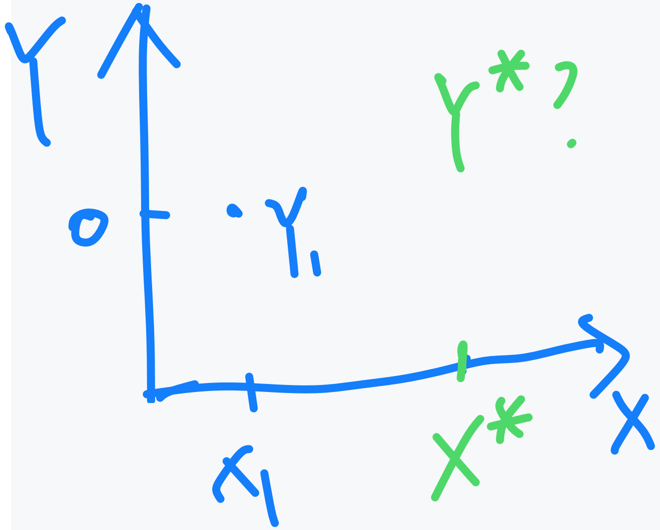
If Y_1 is known, then Y_2 is constrained as $p(Y_2 \mid Y_1)$

Multivariate Gaussian distribution

kernel

which is fully characterized by mean and covariance matrix





obs data $\{Y_i\}$

$$P(Y^*, \{Y_i\}) = P(Y^* | \{Y_i\}) \cdot P(\{Y_i\})$$

$$P(Y^* | \{Y_i\}) = \frac{P(Y^*, \{Y_i\})}{P(\{Y_i\})}$$

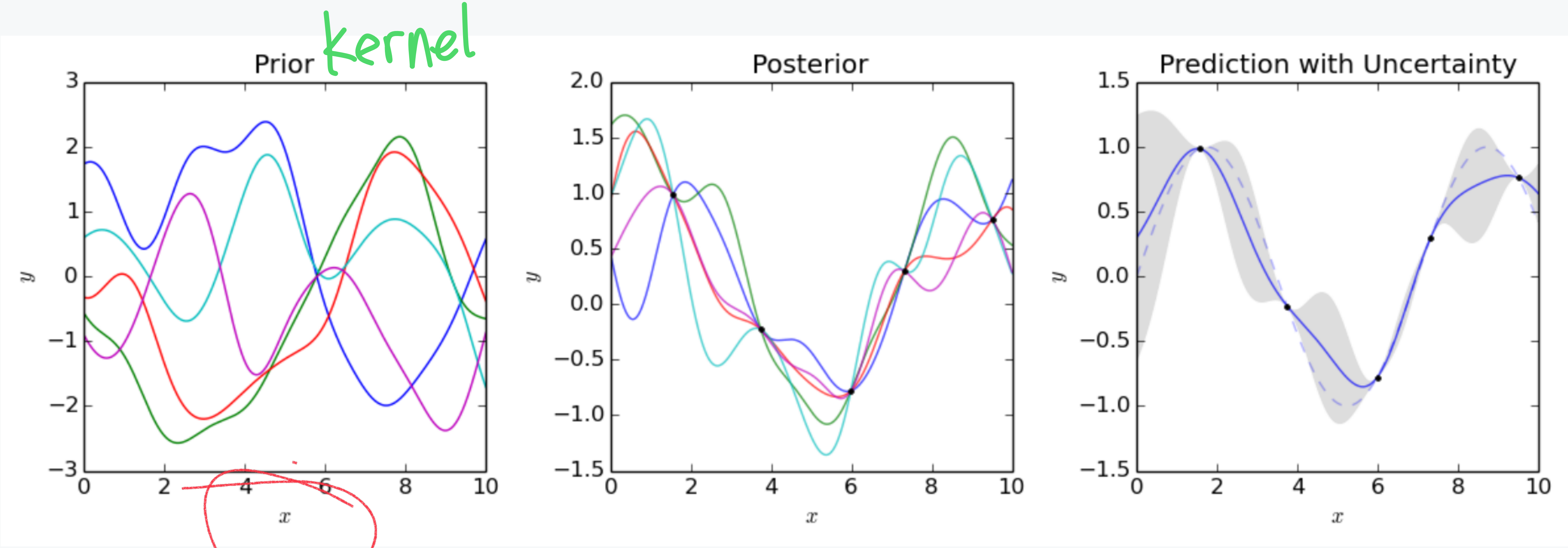
→ Prior

= constant

P(obs data)

If Y_1 is known, then Y^* is constrained as $p(Y^* | Y_1)$

Gaussian Process (GP)



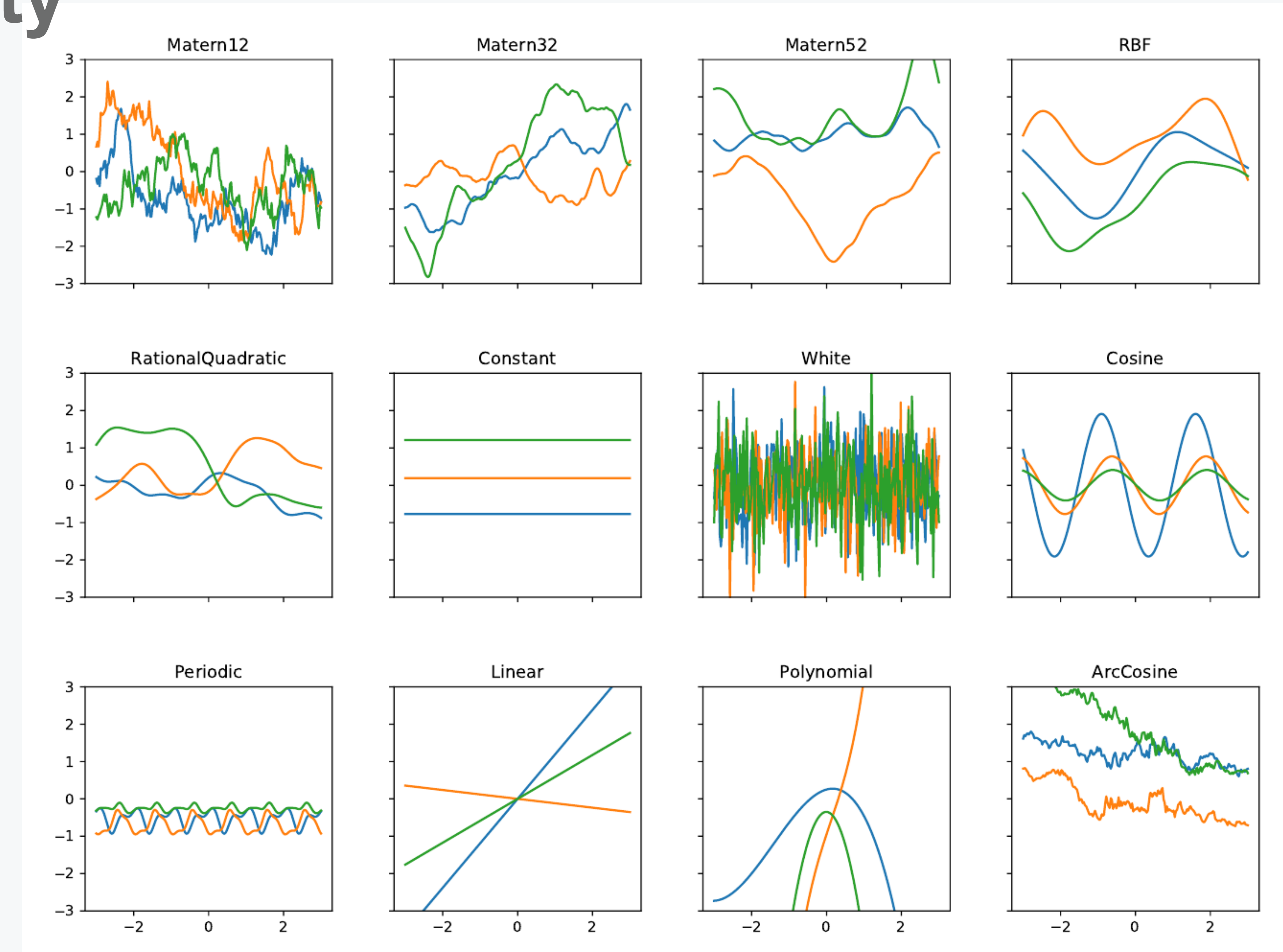
When we know
the mean function m(x)
and the covariance $cov(y_i, y_j)$
between any x_i, x_j
=> A family of functions!

When we further know the y value
at several positions $\{y_i\}$
=> we can predict y at new x*:
$p(y^* \mid \{y_i\})$

The posterior mean and the
uncertainty

Flexibility

Kernel



However, GP is not robust to outliers

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

Astronomy and Computing 36 (2021) 100483

Contents lists available at ScienceDirect

Astronomy and Computing

journal homepage: www.elsevier.com/locate/ascom

ELSEVIER

Full length article

Robust Gaussian process regression based on iterative trimming

Zhao-Zhou Li [a,*], Lu Li [b,c], Zhengyi Shao [b,d]

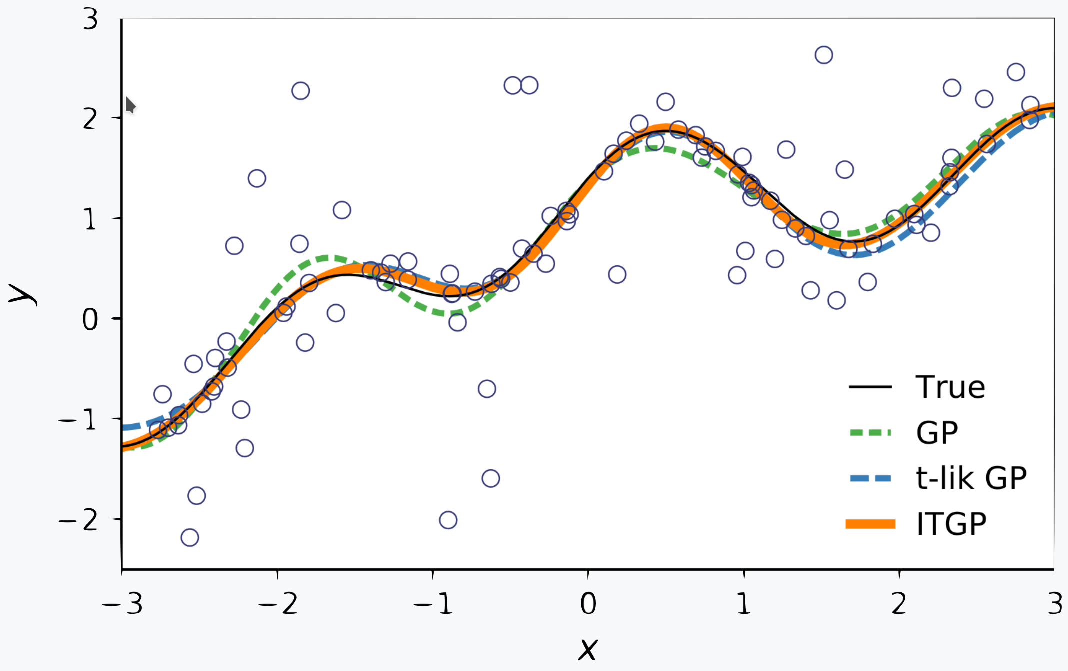
[a] Department of Astronomy, School of Physics and Astronomy, Shanghai Jiao Tong University, 955 Jianchuan Road, Shanghai 200240, China
[b] Key Laboratory for Research in Galaxies and Cosmology, Shanghai Astronomical Observatory, Chinese Academy of Sciences, 80 Nandan Road, Shanghai 200030, China
[c] University of the Chinese Academy of Sciences, No.19A Yuquan Road, Beijing 100049, China
[d] Key Lab for Astrophysics, Shanghai 200234, China

Two solutions:

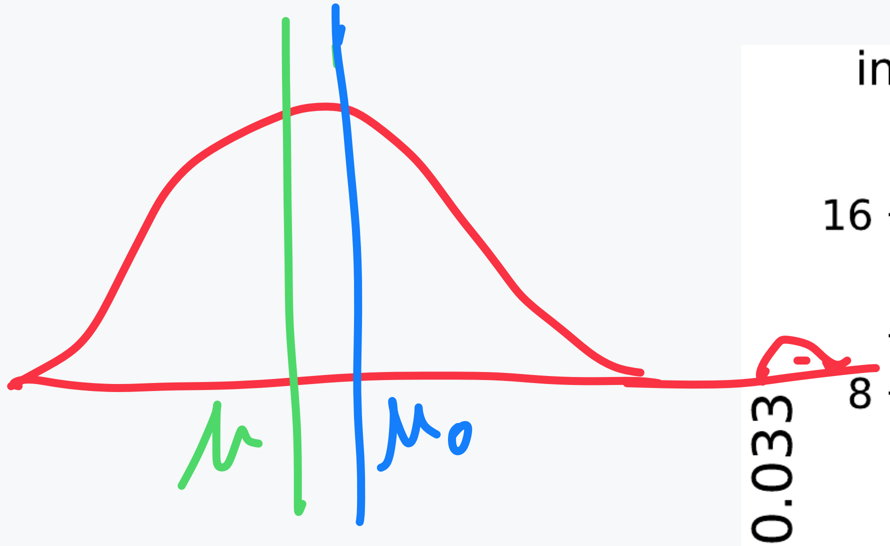
a) lower the significance of poits with large deviation, e.g., using a Students' t distribution for observation model (instead of Gaussian)

b) discard such points
... but in a iterative way

=> Iterative Triming Gaussian Process (ITGP)



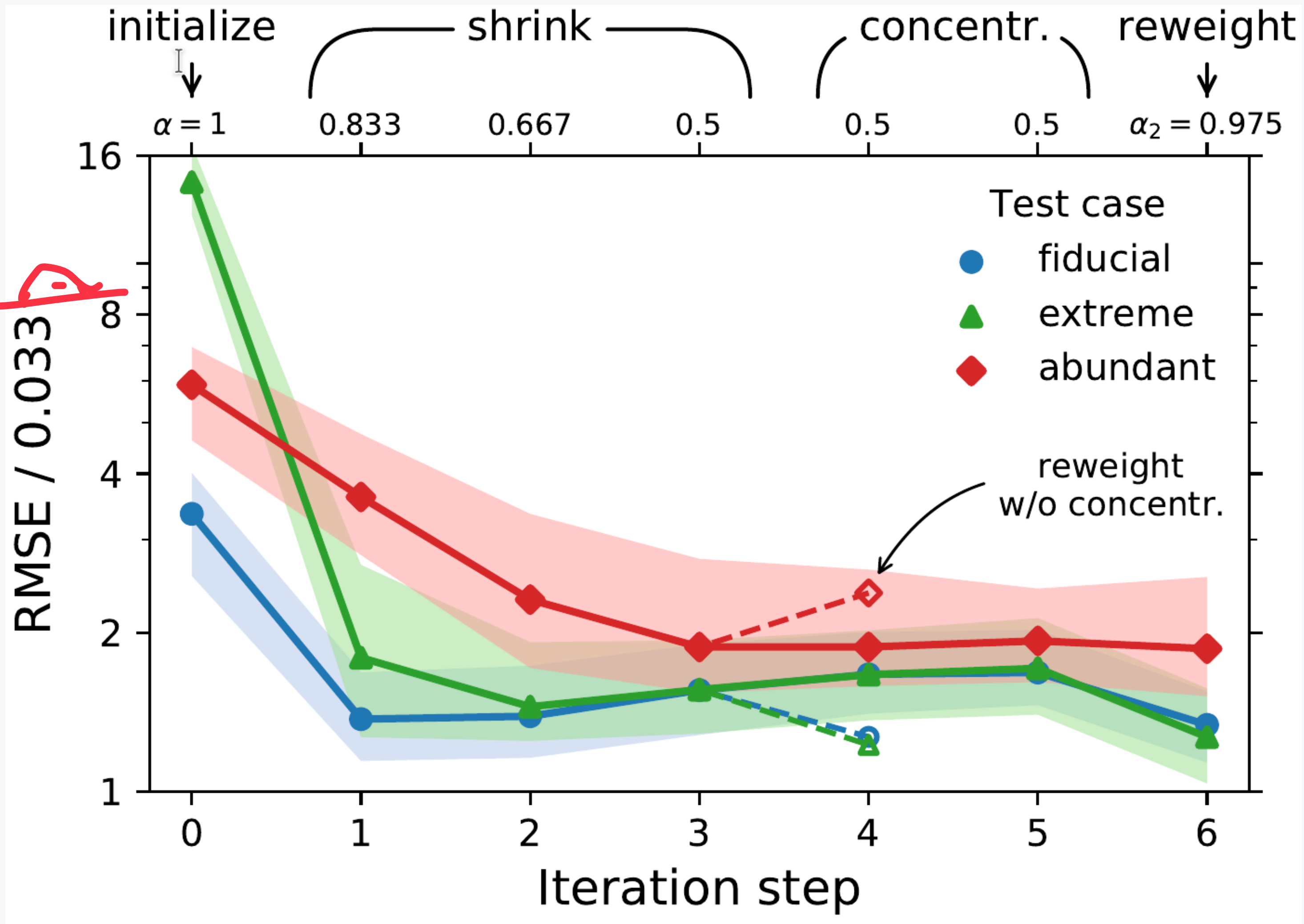
Iterative Trimming Gaussian Process (ITGP)



step 1. trim $3\sigma_0 \to \mu_1, \sigma_1$

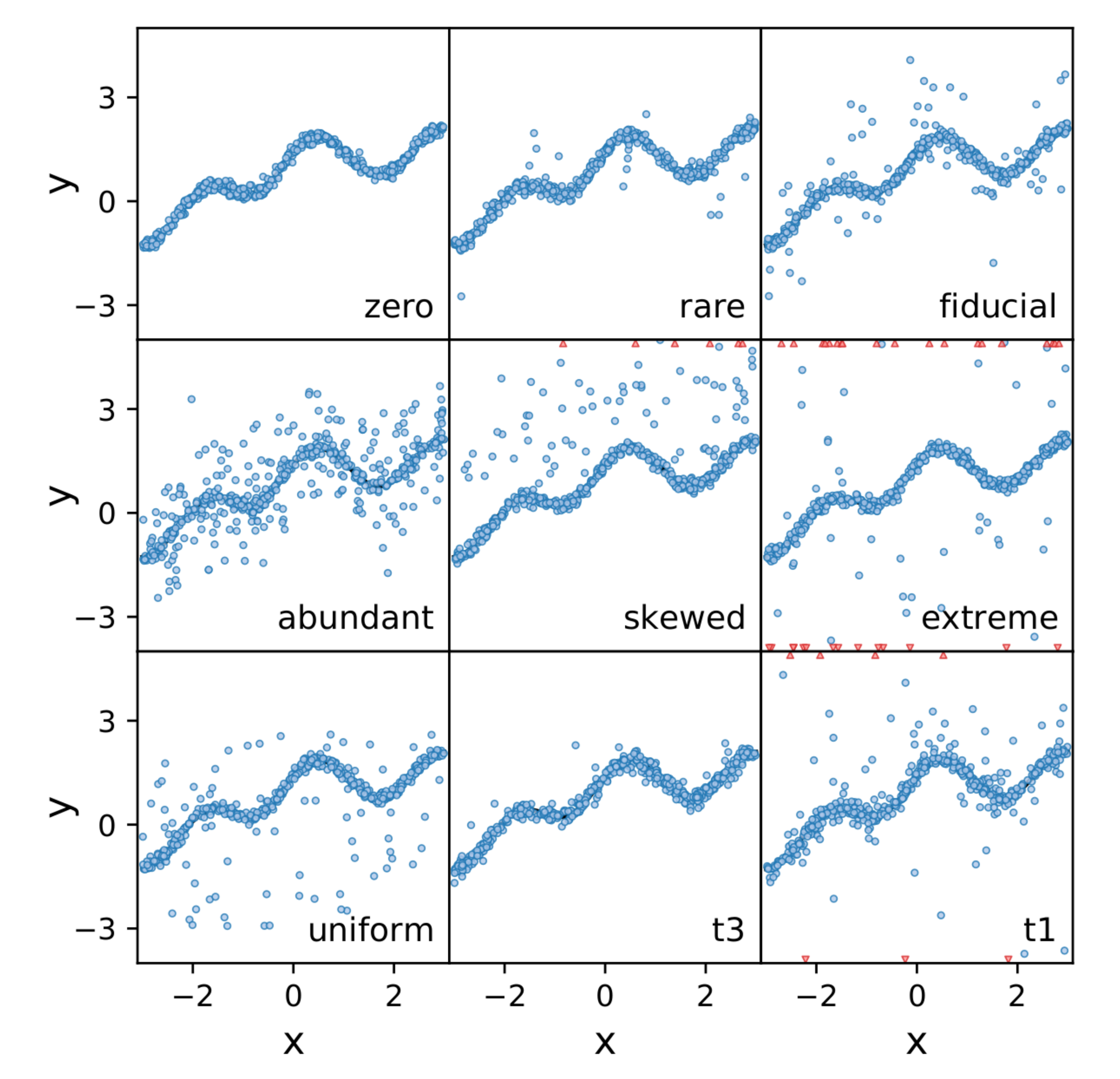
step 2. trim $3\sigma_1 \to \mu_1, \sigma_2$

⋮

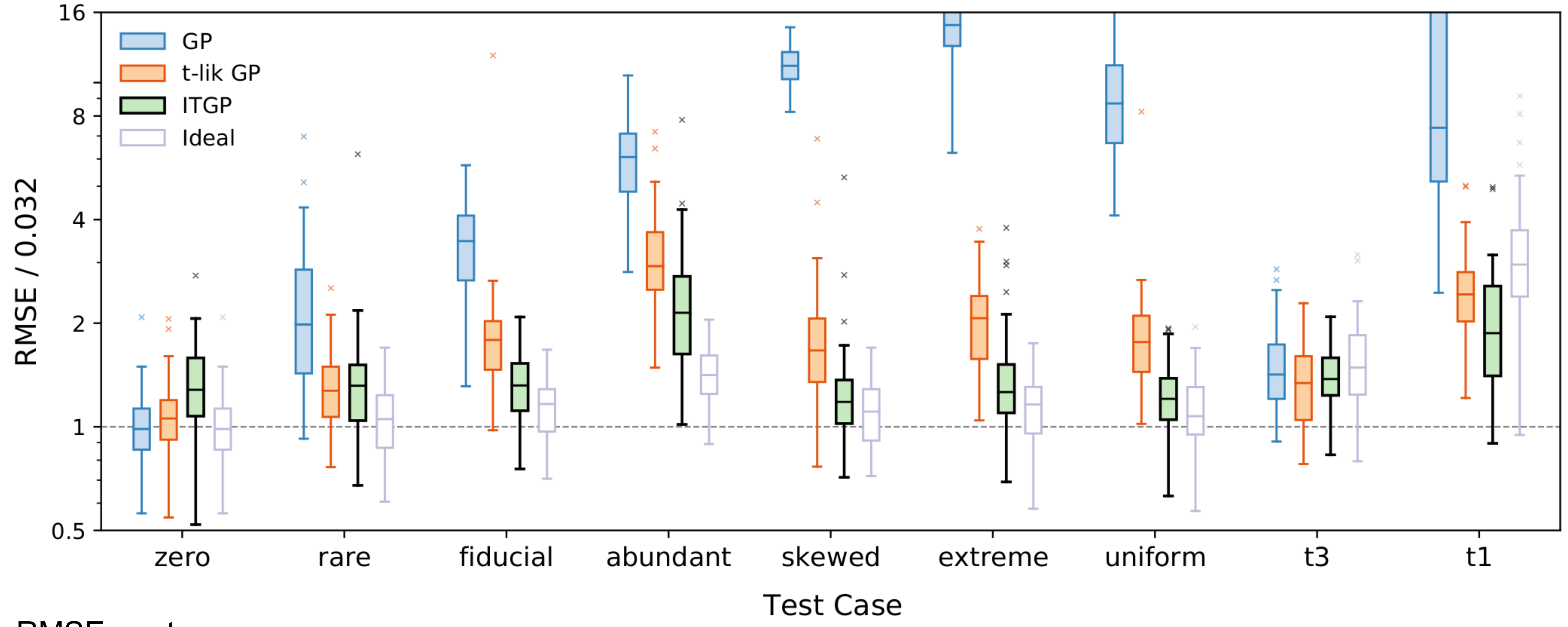


We trim the outliers in an iterative way (ITGP)

Benchmark: test cases

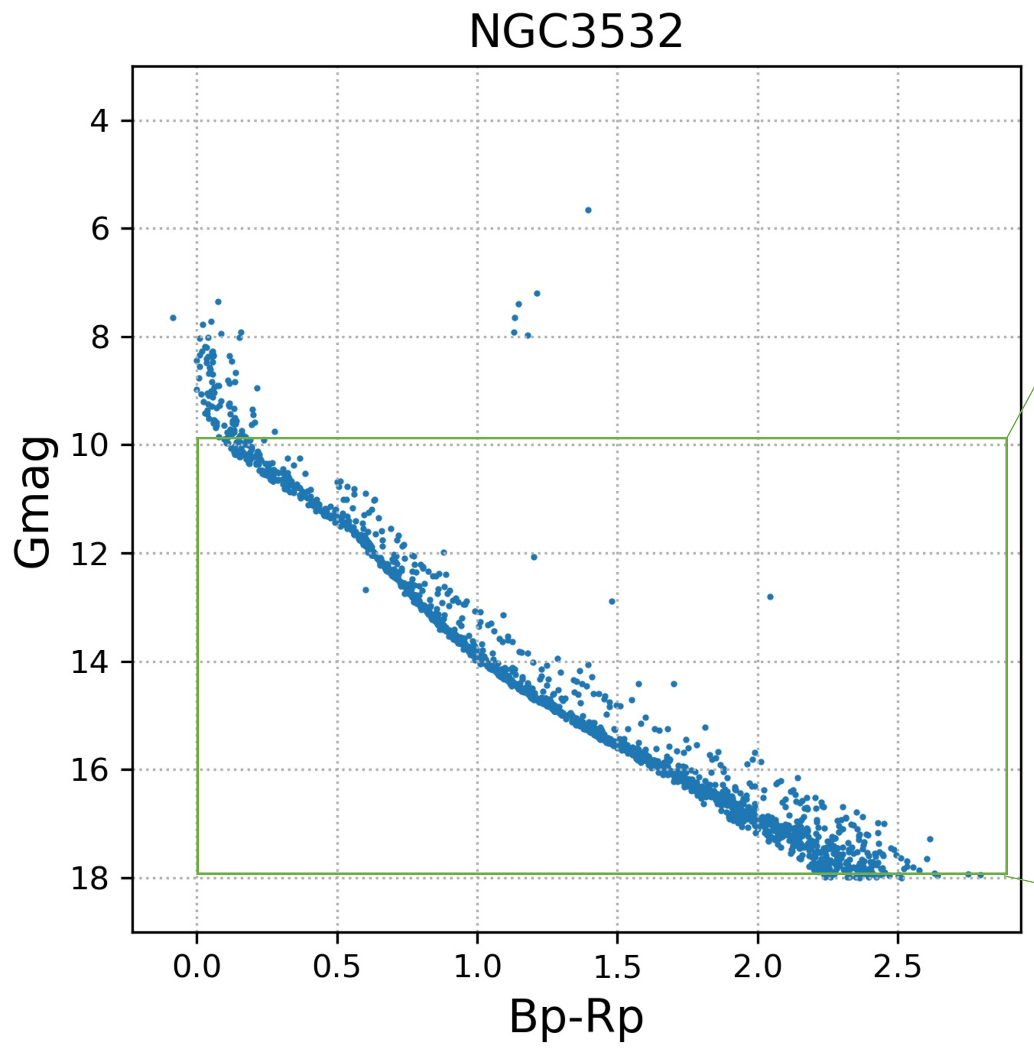


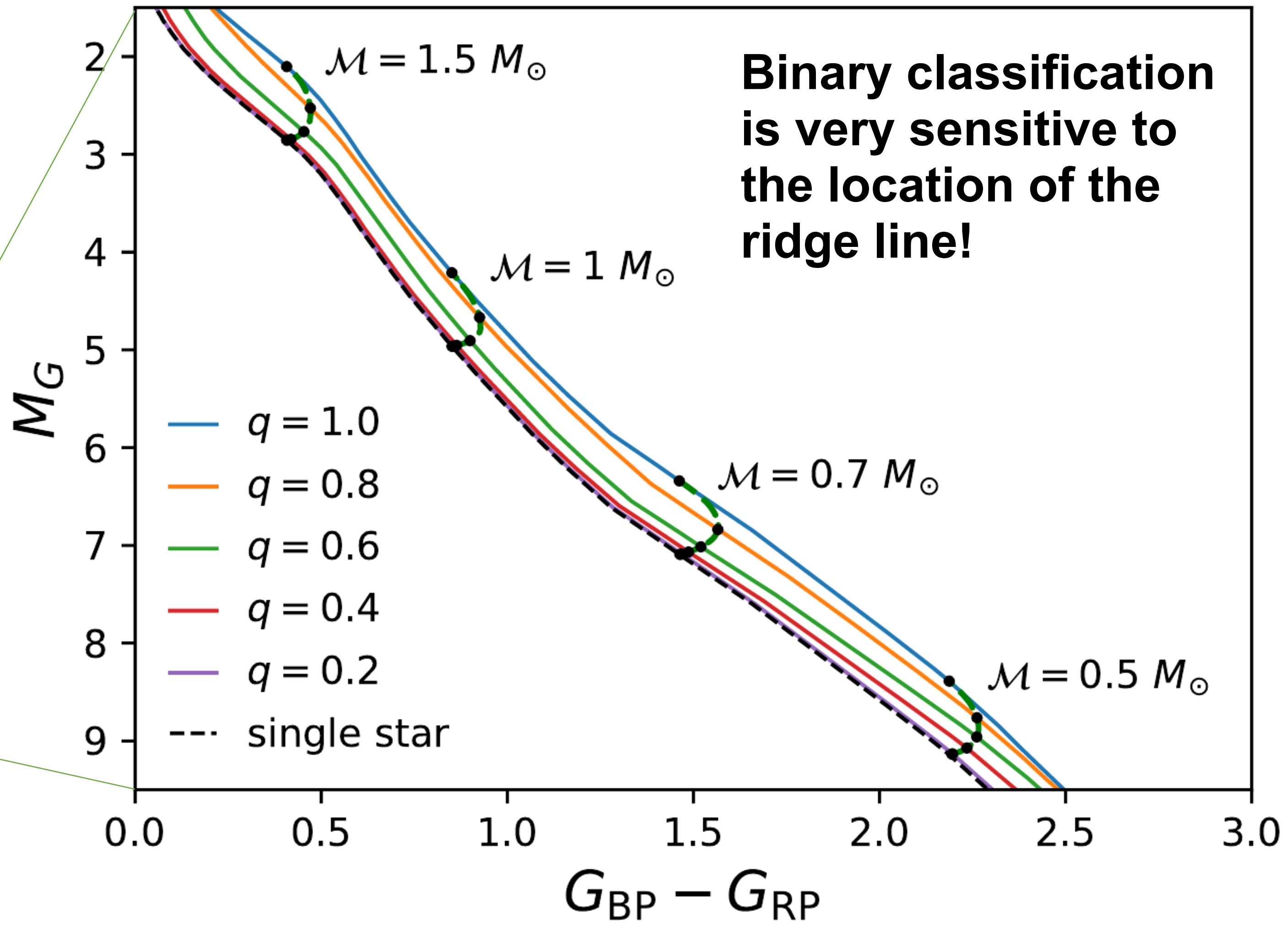
Benchmark



RMSE: root-mean-square error
<(y_pred - y_true)^2>^0.5

Main-sequence ridge line in CMD

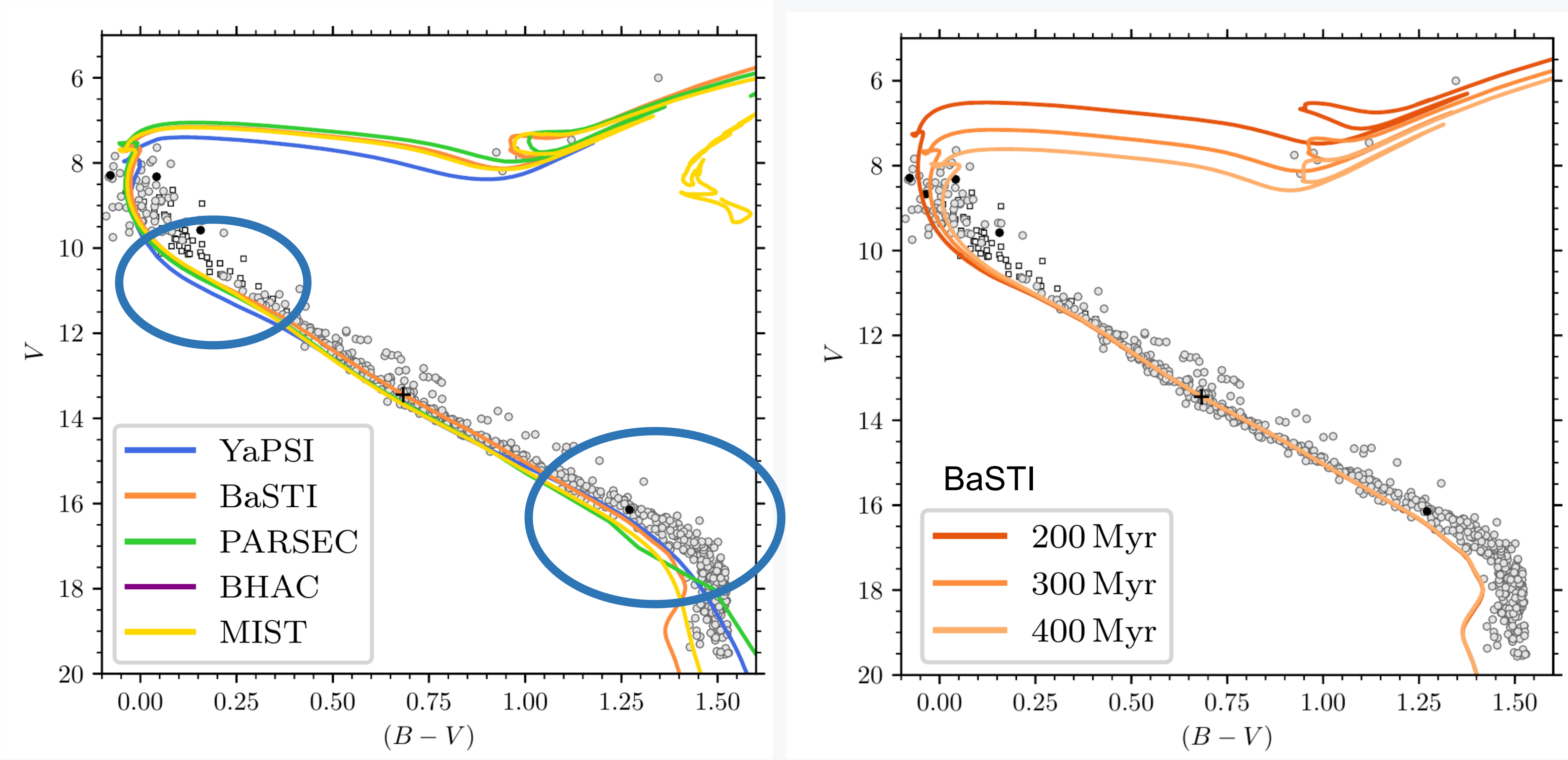




Binary classification is very sensitive to the location of the ridge line!

Lu Li+ 2020

Stellar models are not perfect for observation



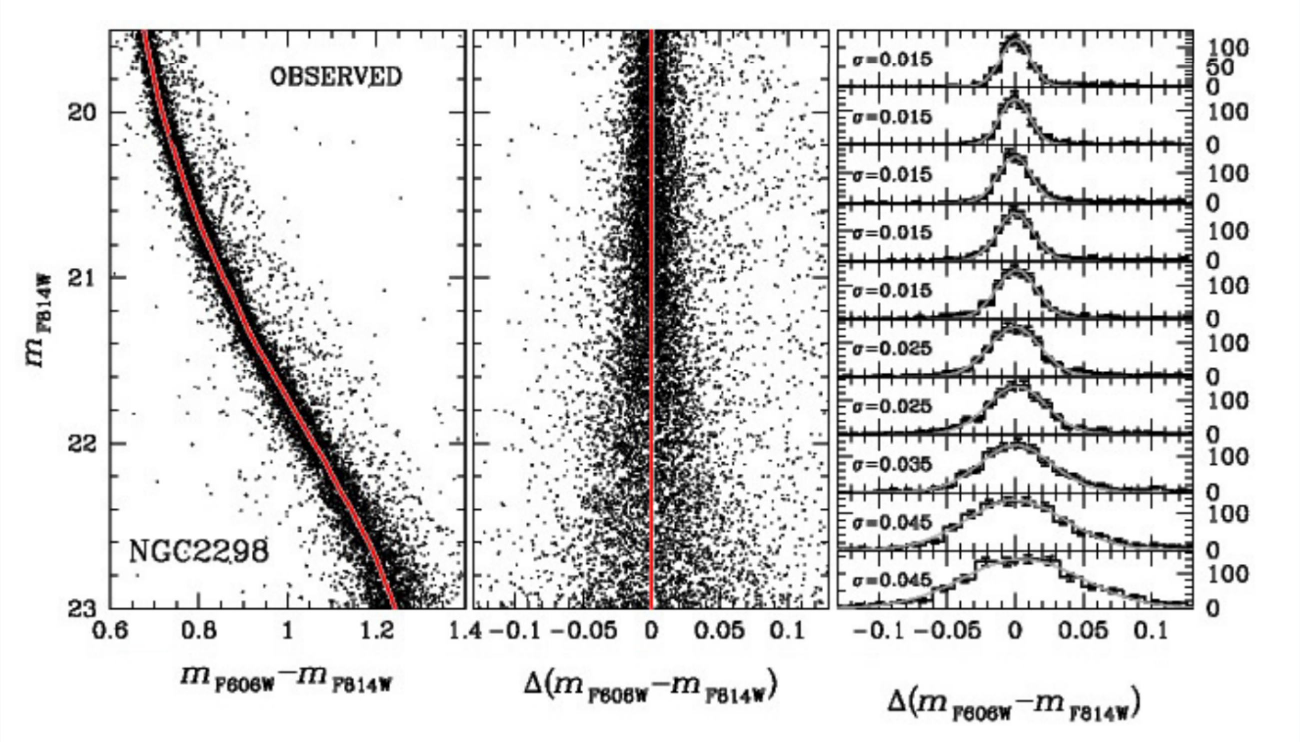
Fritzewski+ 2019

Observed ridge lines of clusters can be used to calibrate stellar models

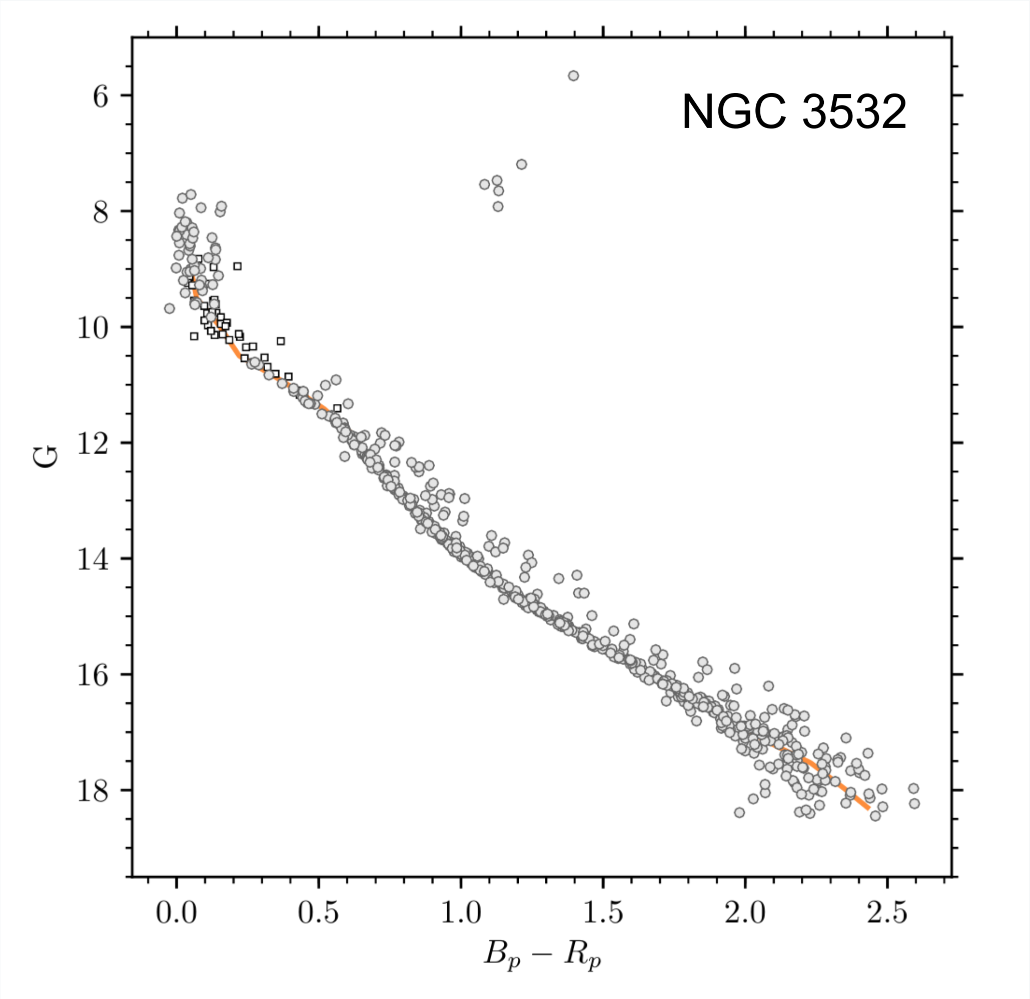
How to let the data itself talk?

Draw the line manually
➢ Arbitrary, lack of reproducibility
➢ Not practical when you have thousands of clusters

Use the peak of histogram of color
➢ Requiring many many member stars (e.g., globular clusters)
➢ Not applicable to open clusters



Milone et al. 2012

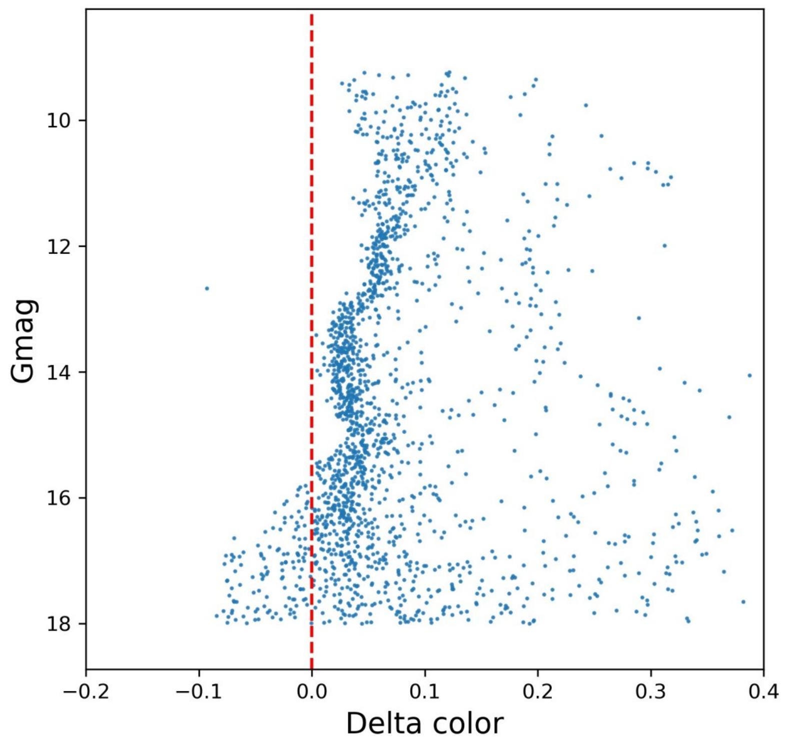


Fritzewski et al. 2019

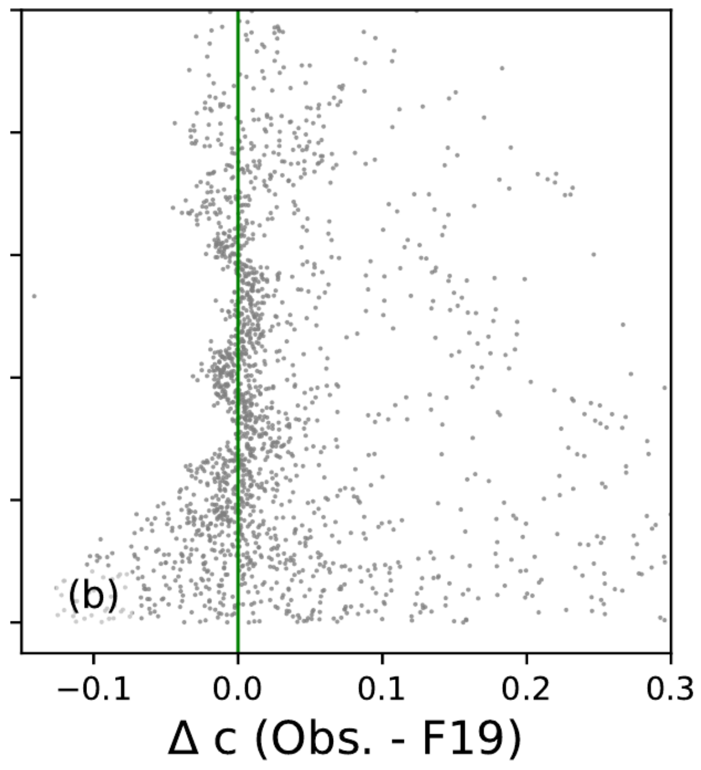
Better solution: Robust GP

Residual color of NGC 3532:

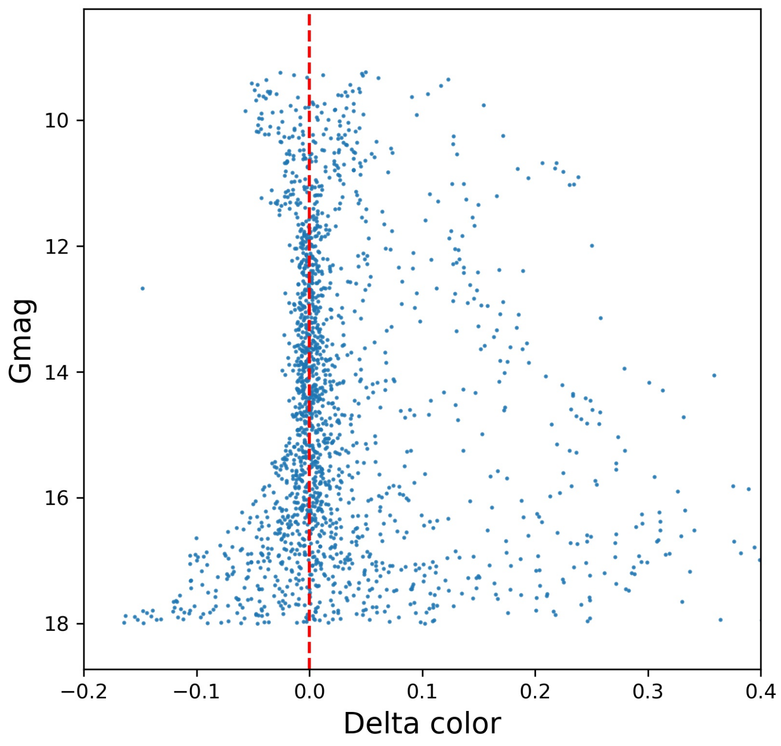
PARSEC (stellar model)



By hand (Fritzewski 19)



Our method (robust GP)



Summery

Gaussian process (GP):

a **non-parametric** and **probabilistic** method

- a non-parametric method, completely data-driven and does not assume any explicit functional form between variables, which is particularly attractive in the big data era.
- provides a Bayesian framework with a natural way of characterizing prior and posterior distributions over functions.
- can naturally handle the noises that are assumed to follow normal distributions.

We can make GP robust by **iterative trimming (ITGP)** outliers

with a practical example in star cluster study

Some sources:

Classic book: Gaussian Processes for Machine Learning, by Rasmussen and Williams 2006

GP tutorial: http://gpss.cc

Python implementation:

for beginner: GPy, scikit-learn

for speed: pytorch, GPflow, celerite

application: 2008.04684

method: 2011.11057

code: https://github.com/syrte/robustgp/