Machine Learning for Regression

Hyungsuk (Tak) Tak

Department of Statistics
Department of Astronomy & Astrophysics
Institute for Computational and Data Sciences
The Pennsylvania State University

Summer School in Astroinformatics II
June 13, 2022

Regression

Regression is an effort to describe the relationship between the response variable Y and p predictor variables (p features) $x_1, x_2, \ldots, x_p$.

Example: Modeling relationship between distance modulus (Y) and redshift for Type Ia Supernovae (x_1) is helpful for understanding different astrophysical theories (e.g., ΛCDM).

Example: Modeling relationship between Y (redshift) and $\{x_1, x_2, x_3, x_4, x_5\}$ (magnitudes in u, g, r, i, z bands) from photometry is essential for LSST and other photometric surveys.

Linear Regression

Linear Regression describes the relationship by a linear model.

Example: For the i-th object,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i, \quad \epsilon_i \sim N(0,\ \sigma^2)$$

redshift

↳ dist. module

Example: For the i-th object,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_5 x_{i5} + \epsilon_i, \quad \epsilon_i \sim N(0,\ \sigma^2)$$

redshift x_{i1} u x_{i2} g x_{i5} z

Unifying matrix notation for any linear regression model:

$$Y_i = \begin{pmatrix} 1 & x_{i1} & x_{i2} & \cdots & x_{ip} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \epsilon_i = x_i^T \beta + \epsilon_i$$

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}_{Y_{n\times 1}} = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}_{X_{n\times(p+1)}} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}_{\beta_{(p+1)\times 1}} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}_{\epsilon_{n\times 1}}$$

$\epsilon \sim N_n(0, \Sigma)$ — a vector of zeros

↳ $\sigma^2 I_n$

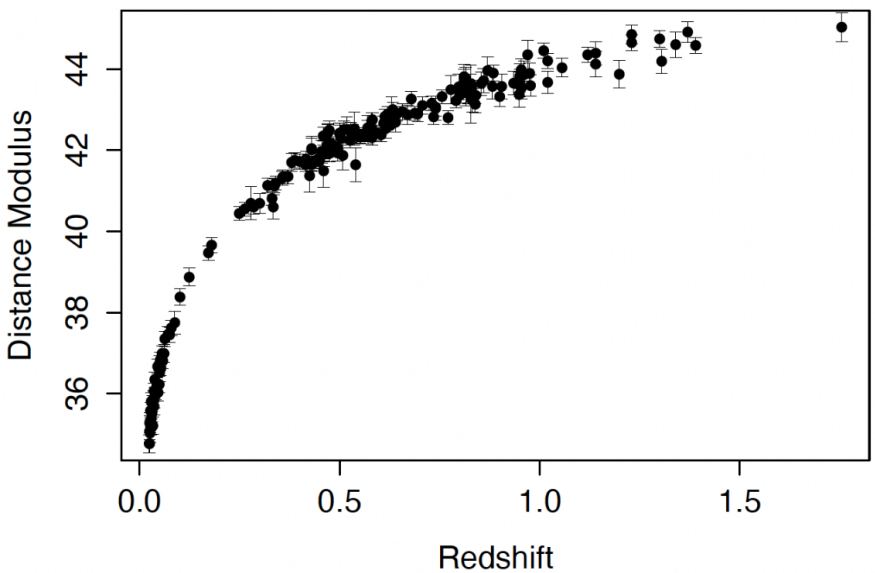
n-dim. multivariate Normal

Linear regression: Meaning of linearity

Linearity in a linear model means that $E(Y)$ is explained by a linear function of the regression coefficients β. That means, a model is linear as long as it is expressed as $E(Y) = X\beta$

Example: Learning the relationship between redshift and distance module can constrain Ω_m and H_0 under the ΛCDM (Riess et al., 2007).

DM



$$DM_i = \beta_0 + \beta_1 Red_i + \beta_2 Red_i^2 + \varepsilon_i$$

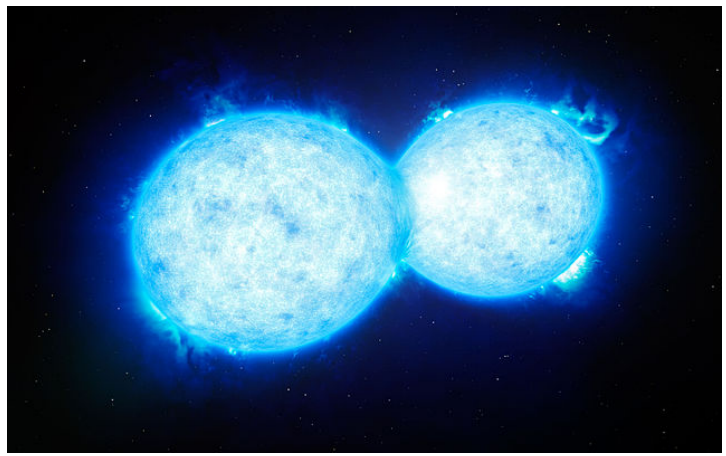
$$Y = X\beta + \varepsilon$$

$$= \begin{pmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 \end{pmatrix} \beta + \varepsilon$$

Thus, a non-linear (quadratic) relationship between DM and redshift is being described by a linear model.

Linear regression: Meaning of linearity (cont.)

Example: The period-luminosity-color relation in early-type contact and near-contact binary stars.



The mathematical formulation of the period-luminosity-color relation (e.g., Pawlak, 2016; Almeida et al., 2015; Rucinski, 2004) is usually stated as

$$M_V = \beta_0 + \beta_1 \log(P) + \beta_2 (V - I)_0,$$

where M_V is the absolute V-band magnitude, P denotes the period and $(V - I)_0$ is the color between the filters V and I.

$= \log \frac{F_V}{F_I}$

Still within a linear model $Y = X\beta + \varepsilon$

Linear Regression: Maximum likelihood

Log-likelihood function of the model parameters, $\theta = (\beta, \Sigma)$, is used to infer the most likelihood values of θ, i.e., maximum likelihood estimates

$$\hat{\theta}_{MLE} = \arg\max_{\theta} \ \ell(\theta), \quad \text{where}$$

$$\ell(\theta) = \ln(L(\theta)) = c - \frac{1}{2}\ln(|\Sigma|) - \frac{1}{2}(y - X\beta)^\top \Sigma^{-1}(y - X\beta)$$

In machine learning, people typically infer θ as values that minimize a quadratic loss function. For example, least square estimates for β are the values that minimize a quadratic loss function:

$$\hat{\beta}_{LS} = \arg\min_{\beta} \ (y - X^\top \beta)^\top \Sigma^{-1}(y - X\beta)$$

Note: The two approaches are equivalent because maximizing $\ell(\theta)$ is equivalent to minimizing $-\ell(\theta)$.

$$\hat{\beta}_{MLE} = \arg\max_{\beta} \ \ell(\theta) = \hat{\beta}_{LS} = \arg\min_{\beta} \ [-\ell(\theta)]$$

Advanced regression modeling in astronomy

1. More general measurement error covariance matrices in a linear model.

Example: In astronomy, (1) most measurement data accompany 'completely known heteroscedastic measurement error uncertainties', and (2) response variables Y are not always independently measured, e.g., light curves.

2. Generalized linear models (non-linear models).

Example: In astronomy, response variables Y are not always Gaussian, e.g., Poisson counts and binary values.

3. Multi-output (multivariate) regression model.

Example: In astronomy, response variables Y can be vector-valued, e.g., multi-band light curves.

More general covariance matrices

Most astronomical data come with completely known heteroscedastic measurement error uncertainties.

The simplest scenario is to assume that X is measured without error (though not the case in astronomy; see Kelly (2007)) but only Y comes with the known heteroscedastic measurement error uncertainties.

Example: A weighted linear regression. Typical astronomical data are

Obs. number	Response Y redshift	Y's 1σ uncertainty	Features X u	g	r	i	z
1	y_1	σ_1	x_{11}	x_{12}	$\cdots$	$\cdots$	x_{15}
2	y_2	σ_2	x_{21}	x_{22}	$\cdots$	$\cdots$	x_{25}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$			$\vdots$
n	y_n	σ_n	x_{n1}	x_{n2}	$\cdots$	$\cdots$	x_{n5}

More general covariance matrices (cont.)

Example (cont.): For the i-th object,

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_5 x_{i5} + \epsilon_i, \quad \epsilon_i \sim N(0, w_i \sigma^2)$$

$w_i \to \frac{1}{\sigma_i^2}$

In matrix notation,

$$Y = X\beta + \varepsilon, \quad \varepsilon \sim N_n(0, \Sigma)$$

$$\Sigma = \sigma^2 \begin{pmatrix} w_1 & & 0 \\ & w_2 & \\ 0 & & \ddots \\ & & & w_n \end{pmatrix}$$

scalar

n-dim. diagonal matrix

Computationally, the model is still manageable because Σ is diagonal, and Σ^{-1} is easy to compute with $O(n)$ cost.

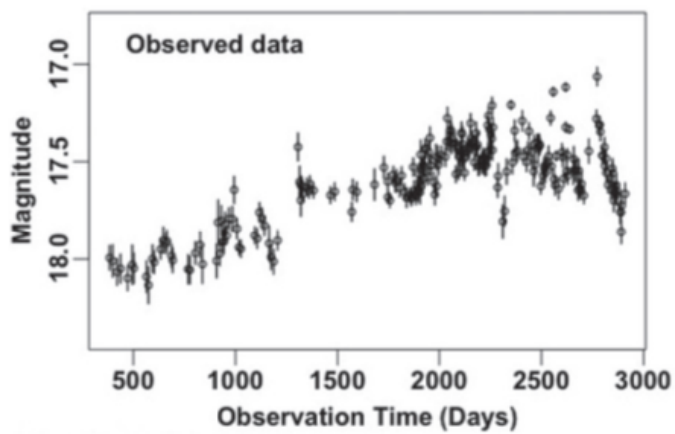
$$= \begin{pmatrix} 1/w_1 & & 0 \\ & 1/w_2 & \\ 0 & & \ddots \\ & & & 1/w_n \end{pmatrix}$$

More general covariance matrices (cont.)

In astronomy, response variables Y aren't always measured independently. Often, they are dependent, e.g., time-dependent structure in light curves.

Example: A Gaussian Process (GP) regression. Typical astronomical data for a single-band light curve are

Obs. number	Obs. time	Response Y mag.	Y's 1σ uncertainty
1	t_1	y_1	σ_1
2	t_2	y_2	σ_2
⋮	⋮	⋮	
n	t_n	y_n	σ_n



More general covariance matrices (cont.)

Example (cont.): The (time-)dependence among $Y = (Y_1, Y_2, \ldots, Y_n)$ is modeled by a non-diagonal covariance matrix Σ.

In matrix notation, $Y = X\beta + \varepsilon = \begin{pmatrix} \beta_0 \\ \beta_0 \\ \vdots \\ \beta_0 \end{pmatrix} + \varepsilon, \quad \varepsilon \sim N_n(0, \Sigma)$

→ overall avg. mag. β_0 (unknown)

$$\Sigma = \begin{pmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \ddots \\ & & & \sigma_n^2 \end{pmatrix} + \begin{pmatrix} \Sigma_{11} & \Sigma_{12} & \cdots & \Sigma_{1n} \\ \Sigma_{21} & \Sigma_{22} & \cdots & \Sigma_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{n1} & \Sigma_{n2} & \cdots & \Sigma_{nn} \end{pmatrix}$$

measurement error component (known)

time-dep. component → $\frac{n(n+1)}{2}$ unknowns too many?

Ex. Matérn ($\frac{1}{2}$) kernel (damped random walk process)

$\Sigma_{ij} = \sigma^2 \exp\left(-\frac{|t_i - t_j|}{\tau}\right)$, now just two unknowns, (σ^2, τ)

The computational cost for computing $\ell(\theta)$ is now prohibitive because Σ^{-1} in $\ell(\theta)$ involves $O(n^3)$ cost.

More general covariance matrices (cont.)

To overcome the daunting computational cost, astronomers have come up with computationally scalable GP models with $O(n^3) \to O(n)$.

Example: Kalman-filtering for CARMA(p, q)-type dependence structure (Kelly et al., 2014). It converts a computation of n-dim. multivariate Gaussian density to a multiplication of n univariate Gaussian densities.

$$L(\beta, \Sigma) = f(y_1, y_2, \ldots, y_n \mid \beta, \Sigma)$$

→ n-dim. MVN density with Σ^{-1}, culprit of $O(n^3)$

$$\overset{\text{KF}}{=} f_1(y_1 \mid \beta, \Sigma)\, f_2(y_2 \mid D_2, \beta, \Sigma) \cdots f_n(y_n \mid D_n, \beta, \Sigma)$$

$$D_i = (y_1, y_2, \ldots, y_{i-1})$$

A product of n univariate Gauss. densities, $O(n)$

Example: Other well-known scalable GP methods in astronomy are (but not limited to) celerite, S+LEAF, Quasi-separable kernels, etc.

Generalized linear models (GLMs)

In astronomy, response variables Y are not always modeled by Gaussian. For example, $Y_1, \ldots, Y_n$ may be integers like Poisson counts or binaries.

Example: Harris et al. (2013) study how the globular cluster population size is correlated somehow with global host galaxy properties. Their catalog data look as follows.

Obs. number	ID	Response Y	Features X		
		N_{GC}	M_V	logMd	logMGC
1	MilkyWay	160	-21.30	9.856	7.66
2	NGC7814	150	-20.18	10.879	7.57
⋮	⋮	⋮	⋮		⋮
256	NGC7768	3000	-23.07	12.142	9.04

Here N_{GC} is a measurement of globular cluster (GC) population size, M_V is the host galaxy visual magnitude in V-band, logMd is dynamical mass of the galaxy on a log scale, and logMGC is total stellar mass contained in the entire GC population of the galaxy on a log scale.

GLMs (cont.)

Can we still use a linear model in this case?

A linear model assumes that response variables Y are Gaussian random variables, meaning that each of Y is assumed to be real-valued on $\mathbb{R}$. But Y here are count data, meaning that each of Y here is integer-valued on $\{0, 1, \ldots\}$. That is, a Gaussian assumption on Y is not quite correct here.

We can still use a linear model, assuming that N_{GC} are Poisson counts, because Poisson counts can be approximated by Gaussian (central limit theorem). But this Gaussian approx. is accurate only for large N_{GC}.

A generalized linear model is a unified framework to relate a wide class of possible values of Y (including non-Gaussian cases) to features X. It provides a way to model N_{GC} as Poisson counts themselves w/o approx.

GLM: Poisson regression

Poisson regression (log-linear model) if $Y_i \in \{0, 1, 2, \ldots\} \sim \text{Pois}(\lambda_i)$

$$\log(\lambda_i) = x_i^\top \beta \quad \rightarrow \quad \lambda_i = \exp(x_i^\top \beta)$$

log is linear

That is,

$$Y_i \sim \text{Pois}(\exp(x_i^\top \beta)).$$

Example:

$$N_{GC,i} \sim \text{Pois}(\ \exp(\beta_0 + \beta_1 M_{V,i} + \beta_2 \log(Md_i) + \beta_3 \log(MGC_i))\ ).$$

Its likelihood function is

$$L(\beta) = \prod_{i=1}^{n} Pr(N_{GC,i} = y_i \mid \beta) = \prod_{i=1}^{n} \frac{\exp(x_i^\top \beta)^{y_i} \exp(-\exp(x_i^\top \beta))}{y_i!}.$$

Pois (λ_i) mass func. is $\frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}$

So, we can conduct a maximum likelihood estimation.

GLM: Poisson Regression (cont.)

R Output:

```
Call:
glm(formula = NGC ~ Mv + logMd + logMGC, family = poisson)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.9820  -0.2255  -0.0857   0.1034   1.0430

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.570616   0.034900 -274.23   <2e-16 ***
Mv           0.139277   0.002972   46.86   <2e-16 ***
logMd       -0.001486   0.004791   -0.31    0.756
logMGC       2.301263   0.004550  505.81   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Computation cost of glm in R: $O(p^3(1+n))$, where n = # of obs. and p = # of features

GLM: Logistic Regression

Binary variables are ubiquitous because it is always possible to make an indicator variable for any event of interest. $Y = \begin{cases} 1, & \text{high-}z \text{ quasars} \\ 0, & \text{otherwise} \end{cases}$

Example: Suppose we have an access to five colors (features) for each of 649,439 objects (20,640 high-z quasars, and 628,799 non-high-z quasars).

Obs. number	Response Y (binary)	Features X					
		ug	*gr*	*ri*	*iz*	*zs*1	*s*1*s*2
1	0	0.295	0.073	···	···	···	0.408
2	0	0.173	-0.035	···	···	···	0.299
⋮	⋮	⋮	⋮				⋮
649439	1	1.687	0.316	···	···	···	-0.161

GLM: Logistic Regression (cont.)

Logistic regression if $Y_i \in \{0, 1\} \sim \text{Bernoulli}(\theta_i)$

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = x_i^\top \beta \quad \rightarrow \quad \theta_i = \frac{1}{1+\exp(-x_i^\top \beta)}$$

logit func. of θ_i

That is,

$$Y_i \sim \text{Bernoulli}\left(\frac{1}{1+\exp(-x_i^\top \beta)}\right).$$

logistic func. of β or Sigmoid func. of β

Its likelihood function is

$$L(\beta) = \prod_{i=1}^{n} Pr(Y_i = y_i \mid \beta) = \prod_{i=1}^{n} \left(\frac{1}{1+\exp(-x_i^\top \beta)}\right)^{y_i} \left(\frac{\exp(-x_i^\top \beta)}{1+\exp(-x_i^\top \beta)}\right)^{1-y_i}.$$

So, we can conduct a maximum likelihood estimation.

GLM: Logistic Regression (cont.)

For simplicity, a subset of the data (size 100,000) is used to fit the model.

R Output:

```
Call:
glm(formula = quasar.indicator ~ ug + gr + ri + iz + zs1 + s1s2
    family = "binomial")

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.4904   0.0171   0.0794   0.1644   4.7351

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.388339   0.058314  92.403   <2e-16 ***
ug          -2.589888   0.035814 -72.315   <2e-16 ***
gr          -0.513882   0.086635  -5.932    3e-09 ***
ri           6.708540   0.160201  41.876   <2e-16 ***
iz           3.817042   0.115019  33.186   <2e-16 ***
zs1         -0.004496   0.030778  -0.146    0.884
s1s2        -3.093086   0.099175 -31.188   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, $O(p^3(1+n))$

Multi-output regression

In astronomy, response variables Y are sometimes vector-valued.

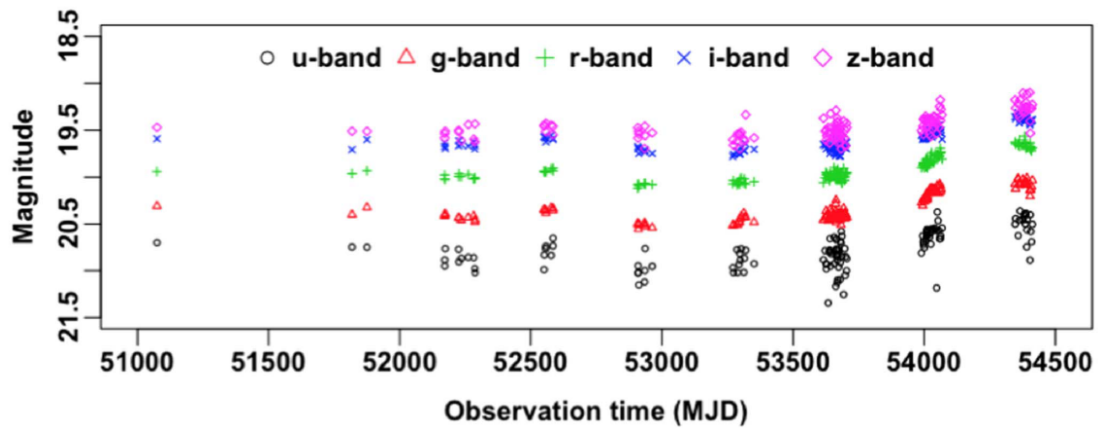
Example: A multi-output (vector-valued) Gaussian process regression. Typical astronomical data for multi-band light curves are

Obs. number	Obs. time	g-mag.	1σ error
1	t_1^g	y_1^g	σ_1^g
2	t_2^g	y_2^g	σ_2^g
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_g	$t_{n_g}^g$	$y_{n_g}^g$	$\sigma_{n_g}^g$

$\cdots$

Obs. number	Obs. time	r-mag.	1σ error
1	t_1^r	y_1^r	σ_1^r
2	t_2^r	y_2^r	σ_2^r
$\vdots$	$\vdots$	$\vdots$	$\vdots$
n_r	$t_{n_r}^r$	$y_{n_r}^r$	$\sigma_{n_r}^r$

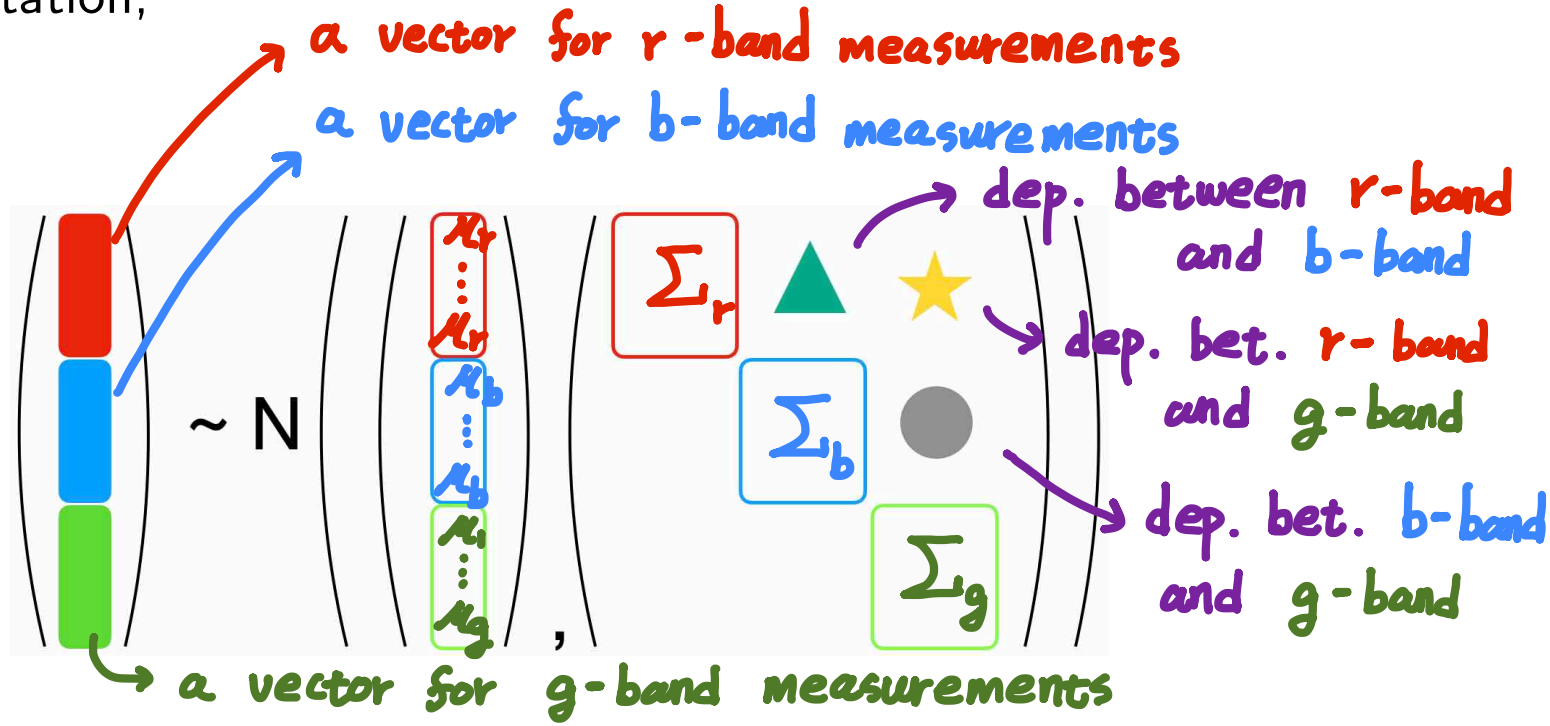
Five-band light curves of quasar 3078106 from a catalog of SDSS S82.



Multi-output regression (cont.)

Example (cont.): Time- and band-dependance among $Y = (Y_r, Y_b, Y_g)$ is now modeled by a really big non-diagonal covariance matrix Σ.

In matrix notation,



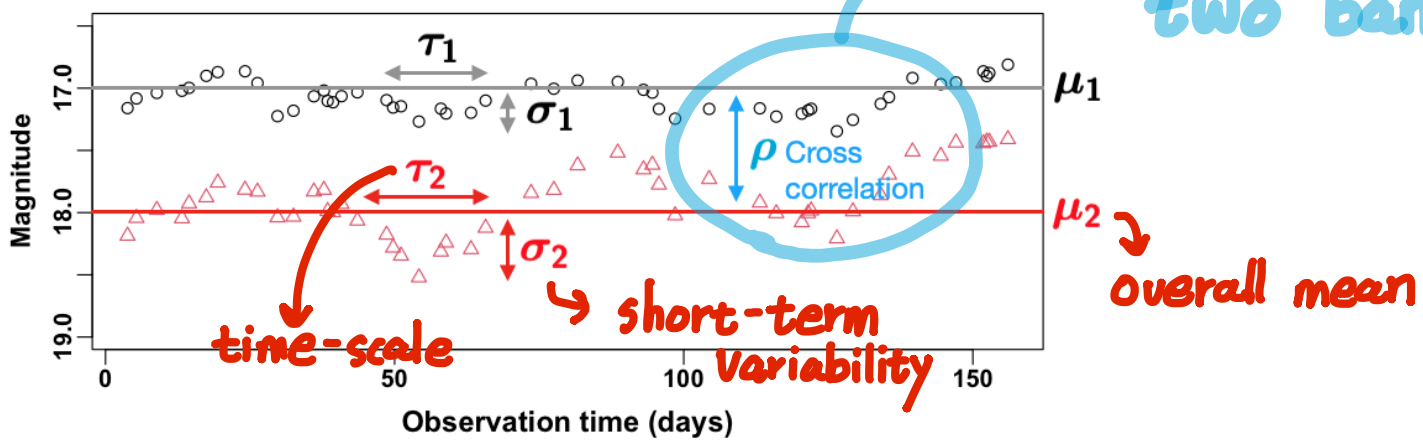
The computational cost for computing $\ell(\theta)$ is now even more prohibitive because Σ^{-1} in $\ell(\theta)$ involves $O((n_r + n_b + n_g)^3)$ cost.

Multi-output regression (cont.)

To overcome the daunting computational cost, astronomers have come up with computationally scalable Gaussian process models with

$$O((n_1 + n_2 + \cdots + n_5)^3) \to O(n_1 + n_2 + \cdots + n_5).$$

Example: Kalman-filtering for vectorized CARMA$(1, 0)$-type dependence structure (Hu and Tak, 2020).



θ, a vector of length 7 for two-band data
" 12 for three-band data