

# INCORPORATING MEASUREMENT ERROR IN ASTRONOMICAL OBJECT CLASSIFICATION

Hyungsuk (Tak) Tak

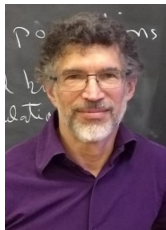
Department of Statistics  
Department of Astronomy & Astrophysics  
Institute for Computational & Data Sciences  
The Pennsylvania State University

Oct 27, 2022

# COLLABORATORS



Sarah Shy (Microsoft)



Eric Feigelson (Astro)



John Timlin  
(Lockheed Martin)

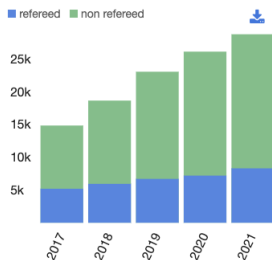


Jogesh Babu (Stat)

# CLASSIFICATION IN ASTRONOMY

Classification methods have played an important role in astronomy for centuries and continue to be essential tools in modern astronomy.

In the NASA Astrophysics Data System, the keyword 'classification' appeared in over 5,000 refereed astronomy/physics papers annually for the last 5 consecutive years (2017–2021).



Year	Article Count	Ref Count
2017	14894	5208
2018	18716	5958
2019	23138	6713
2020	26216	7222
2021	28879	8333

# MEASUREMENT ERROR

Many classification methods have publicly available code implementations such as random forest, neural network, Gaussian processes, etc.

However, astronomical data for classification (right) are not the same as those in a classification textbook (left) due to **measurement error**.

	$x$	$y$	
Labeled	$x_1^{\text{obs}}$	$y_1$	Training
	$x_2^{\text{obs}}$	$y_2$	
	$\vdots$	$\vdots$	
	$x_m^{\text{obs}}$	$y_m$	
Unlabeled	$x_{m+1}^{\text{obs}}$		
	$\vdots$		
	$x_n^{\text{obs}}$		

	$x$	$\sigma$	$y$	
Labeled	$x_1^{\text{obs}}$	$\sigma_1$	$y_1$	Training
	$x_2^{\text{obs}}$	$\sigma_2$	$y_2$	
	$\vdots$	$\vdots$	$\vdots$	
	$x_m^{\text{obs}}$	$\sigma_m$	$y_m$	
Unlabeled	$x_{m+1}^{\text{obs}}$	$\sigma_{m+1}$		
	$\vdots$	$\vdots$		
	$x_n^{\text{obs}}$	$\sigma_n$		

Here,  $x_i^{\text{obs}} = (x_{i1}^{\text{obs}}, x_{i2}^{\text{obs}}, \dots, x_{ip}^{\text{obs}})^{\top}$  and  $\sigma_i = (\sigma_{i1}, \sigma_{i2}, \dots, \sigma_{ip})^{\top}$  obtained by careful calibration of the instrument and examination of source-free regions of the image or spectrum.

## MEASUREMENT ERROR (CONT.)

In astronomy, a Gaussian measurement error assumption dates back to Eddington (1913).

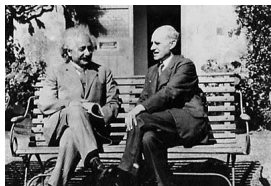


Image Credit: Great War Films.

For a measurement of the  $j$ -th property (e.g., color, brightness, redshift) of the  $i$ -th object,

$$x_{ij}^{\text{obs}} = x_{ij}^{\text{true}} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_{ij}^2),$$

# MEASUREMENT ERROR IN ASTRO. CLASSIFICATION

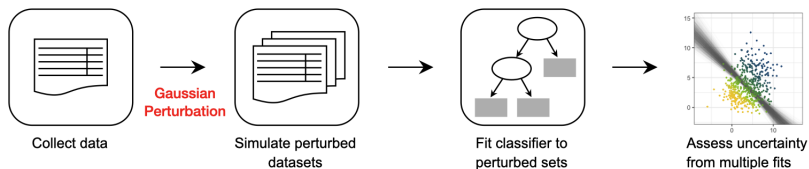
In astronomical object classification, standard classification methods are often adopted by **discarding the uncertainty columns** (e.g., Timlin et al., 2018).

Bovy et al. (2011, 2012) proposed Bayesian methods to account for measurement error in astronomical classification (but limited to quasar classification).

Could we develop a general framework that enables standard classification methods to account for measurement error in astronomical data?

# GAUSSIAN PERTURBATION

**Gaussian perturbation** is a way to replicate data sets by perturbing the observed data with the Gaussian measurement errors (like bootstrapping).



For example, a linear support vector machine is fit on each perturbed set, producing a single decision boundary for each set.

A collection of decision boundaries from multiple fits → **a decision band**.

## GAUSSIAN PERTURBATION (CONT.)

Gaussian perturbation simulates replicates by randomly generating

$$x_{ij}^{\text{rep}} \sim N(x_{ij}^{\text{obs}}, 2\sigma_{ij}^2).$$

This is a **posterior predictive distribution** of a two-level Gaussian hierarchical model. Specifically, from the Gaussian error model

$$x_{ij}^{\text{obs}} | x_{ij}^{\text{true}} \sim N(x_{ij}^{\text{true}}, \sigma_{ij}^2),$$

we assume a flat prior on each unknown true value  $x_{ij}^{\text{true}}, h(x_{ij}^{\text{true}}) \propto 1$ . Then, the posterior distribution of  $x_{ij}^{\text{true}}$  is

$$x_{ij}^{\text{true}} | x_{ij}^{\text{obs}} \sim N(x_{ij}^{\text{obs}}, \sigma_{ij}^2),$$

and the resulting posterior predictive distribution is

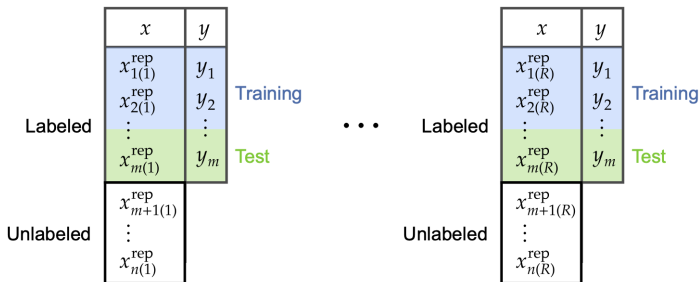
$$q(x_{ij}^{\text{rep}} | x_{ij}^{\text{obs}}) = \int f(x_{ij}^{\text{rep}} | x_{ij}^{\text{true}}) \pi(x_{ij}^{\text{true}} | x_{ij}^{\text{obs}}) dx_{ij}^{\text{true}} = N(x_{ij}^{\text{rep}} | x_{ij}^{\text{obs}}, 2\sigma_{ij}^2).$$



# GAUSSIAN PERTURBATION (CONT.)

The measurement error uncertainty is used **only once** for producing perturbed data sets and is never used anymore.

The resulting replicates don't have columns for measurement uncertainty.



Now, each replicate is in the format shown in a classification textbook.  
Thus, any standard classification methods can be fit on each replicate set.

# CLASSIFICATION VIA GAUSSIAN PERTURBATION

The resulting variation of any quantity of interest from multiple fits forms its posterior predictive distribution (Gelman et al., 2013, Chap. 6).

- ▶ Let  $C$  be a classification operator, mapping a data set to a quantity for classification summary,  $\theta$ , such as classification accuracy, predicted labels, etc.
- ▶ Let us define  $\theta_{(i)} = C(X_{(i)}^{\text{rep}})$  as the quantity of interest obtained by applying the classification operator  $C$  to the  $i$ -th replicate data set.

---

**input:** perturbed data sets  $X_{(1)}^{\text{rep}}, \dots, X_{(R)}^{\text{rep}}$ ,  
classification algorithm  $C$ , metric  $\theta$

**result:** A sample,  $\{\theta_{(1)}, \dots, \theta_{(R)}\}$ , of size  $R$  from the  
posterior predictive distribution of  $\theta$

**for**  $r = 1 \dots R$  **do**  
    | Fit classifier  $C$  to  $X_{(r)}^{\text{rep}}$   
    | Calculate metric of interest  $\theta_{(r)}$   
**end**

---

- ▶ Then, the values of  $\theta_{(i)}$ 's represent the posterior predictive distribution of the quantity of interest.

# SIMULATION STUDY: DATA

Let us consider the following simulation setting with two features.

1. First simulate 200 features each with length 2, and set true labels.

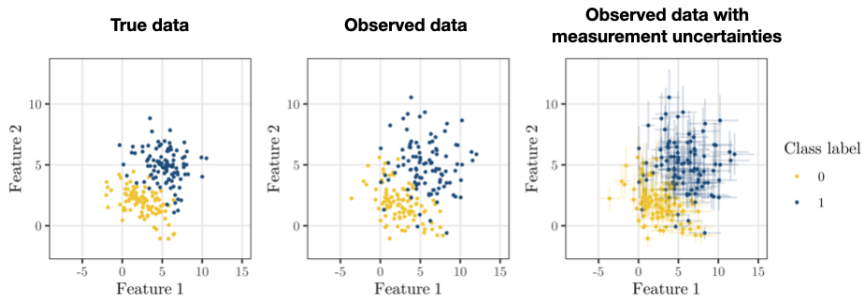
$$\begin{aligned}x_1^{\text{true}}, \dots, x_{100}^{\text{true}} &\sim \mathcal{N}\left(\begin{bmatrix} 5 \\ 5 \end{bmatrix}, \begin{bmatrix} 3 & -0.5 \\ -0.5 & 2 \end{bmatrix}\right) & y_i = 0, \text{ for } i = 1, \dots, 100 \\x_{101}^{\text{true}}, \dots, x_{200}^{\text{true}} &\sim \mathcal{N}\left(\begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 4 & -1 \\ -1 & 1 \end{bmatrix}\right) & y_i = 1, \text{ for } i = 101, \dots, 200\end{aligned}$$

2. Next generate 200 observed data by

$$x_{ij}^{\text{obs}} = x_{ij}^{\text{true}} + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma_{ij}^2).$$

where  $\sigma_{ij}^2 = |x_{ij}^{\text{true}}|/2$  to be realistic.

# SIMULATION STUDY: DATA (CONT.)

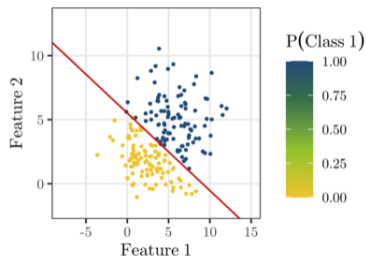


Next, Gaussian perturbation will generate 500 replicate data sets using the data in the 3rd panel.

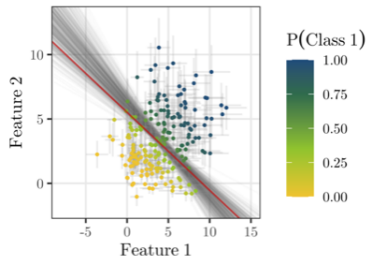
Both linear SVM (a hard classifier) and random forest (a soft classifier) will be fit on each of the 500 replicate sets.

# SIMULATION STUDY: SVM

**Support vector machine  
without measurement uncertainty**



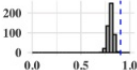
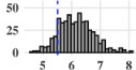
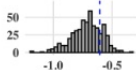
**Support vector machine  
with measurement uncertainty**



The gray decision band (accounting for measurement error) encompasses the red decision band (obtained without measurement error).

The label prediction probabilities show a yellow-green-blue color gradient (softening a hard classification).

# SIMULATION STUDY: SVM (CONT.)

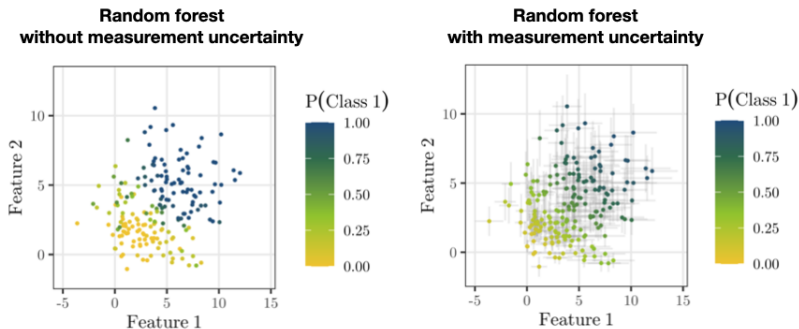
Method	Classification accuracy	SVM decision boundary	
		Intercept	Slope
SVM without measurement error	0.91	5.51	-0.60
SVM with measurement error	$0.80 \pm 0.03$	$6.15 \pm 0.60$	$-0.71 \pm 0.13$
Posterior predictive distribution			

The variation in (cross-validated) classification accuracy stems from the variation in the SVM decision boundary caused by measurement error.

Over-confidence about our measurements (ignoring measurement error) can lead to exaggerated results and potential bias in classification.

The lower accuracy under the Gaussian perturbation is natural as the measurement error has further blurred the separation between classes.

# SIMULATION STUDY: RANDOM FOREST



The yellow-green-blue color gradient becomes more blurred (less extreme) overall due to measurement error uncertainty.

The cross-validated classification accuracy is 0.87 without measurement error, and  $0.77 \pm 0.03$  with measurement error.

# REAL DATA: HIGH-REDSHIFT QUASARS

Let us identify high-redshift quasar candidates ( $2.9 \leq \text{redshift} \leq 5.1$ ) from a catalog data set merged from multiple sources (SDSS, Spitzer IRAC, Spitzer-HETDEX).

	$x$	$\sigma$	$y$	
Labeled	$x_1^{\text{obs}}$	$\sigma_1$	$y_1$	Training
	$x_2^{\text{obs}}$	$\sigma_2$	$y_2$	
	$\vdots$	$\vdots$	$\vdots$	
	$x_m^{\text{obs}}$	$\sigma_m$	$y_m$	
Unlabeled	$x_{m+1}^{\text{obs}}$	$\sigma_{m+1}$		Test
	$\vdots$	$\vdots$		
	$x_n^{\text{obs}}$	$\sigma_n$		

**649,439 labeled objects**  
– 5,487 high-redshift quasars (<1%)  
– 643,952 anything else (stars, galaxies, unidentified quasars)

**1,862,968 unlabeled objects**

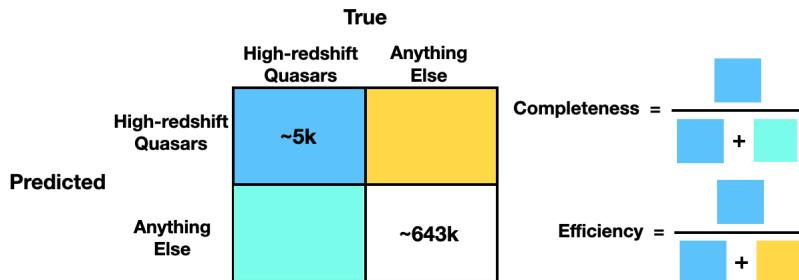
Each observation is composed of 6 colors ( $ug, gr, ri, iz, zs1, s1s2$ ),

$$x_i^{\text{obs}} = \{x_{i1}^{\text{obs}}, x_{i2}^{\text{obs}}, \dots, x_{i6}^{\text{obs}}\}.$$



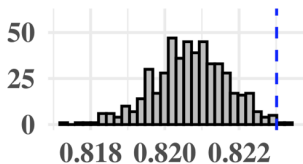
# REAL DATA: CLASSIFICATION ACCURACY

Considering substantial class imbalance, we use two accuracy measures, **completeness** and **efficiency** (Timlin et al., 2018).



# REAL DATA: CLASSIFICATION ACCURACY (CONT.)

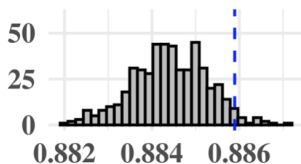
**Posterior predictive distribution  
of completeness**



**Completeness with  
measurement error:  
0.821 +/- 0.001**

**Completeness without  
measurement error:  
0.823**

**Posterior predictive distribution  
of efficiency**



**Efficiency with  
measurement error:  
0.8840 +/- 0.0008**

**Efficiency without  
measurement error:  
0.8860**

Is Gaussian perturbation still useful, considering that both completeness and efficiency are quite similar?

# REAL DATA: PREDICTING UNLABELED OBJECTS

		Without measurement error	
		High-redshift Quasars	Anything Else
With measurement error	High-redshift Quasars	8,701	936
	Anything Else	3,146	1,850,185

936 objects are predicted as AE (Anything Else) without measurement error, but as high-redshift quasars with measurement error.

Potential new candidates (~900 needles) might have been buried in the AE group (haystack of ~2M), without considering measurement error.

3,146 objects are predicted as high-redshift quasars without measurement error, but as AE with measurement error.

Potential misclassification that might have wasted telescope's substantial time and effort, without considering measurement errors.

# DISCUSSION

1. Why Bayesian posterior predictive distribution?

A naive approach is to sample  $x_{ij}^{\text{rep}} \sim N(x_{ij}^{\text{obs}}, \sigma_{ij}^2)$  (Ball+, 2007).

But, this is the post. distribution of  $(x_{ij}^{\text{true}} | x_{ij}^{\text{obs}})$  and  $x_{ij}^{\text{true}} \neq x_{ij}^{\text{rep}}$ .

2. The number of perturbed data sets.

200 replicates are recommended (Efron and Tibshirani, 1994, p48)

3. Additional computational cost linearly increases in terms of the number of perturbed sets.

4. Correlated measurement error.

Possible by modeling correlations (future work).

5. Gaussian perturbation beyond classification, e.g., clustering (unsupervised learning) (future work).

6. Limitations.

Limited to 'Gaussian' measurement error.