Real-Time Light Curve Classification

Dan Cervone

CHASC

October 2, 2012

Introduction

Scientists are interested in studying variable light sources for a number of reasons, including making inferences about the distribution of dark matter and evolution of the universe.

- Number of observable sources vastly outscales resources for observation.
- Astronomers seek to maximize the information (per unit time) given from their limited resources.
- Don't want to waste time and imagery on sources that don't give us new or useful information.

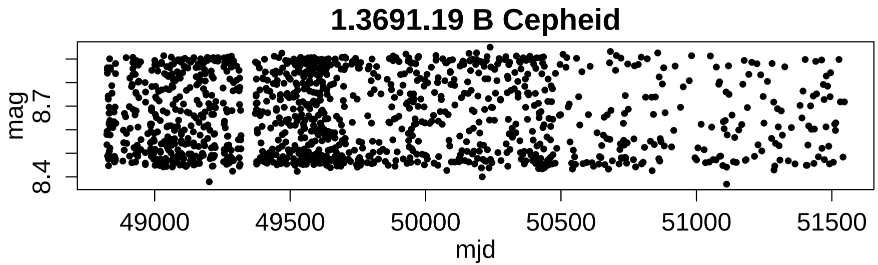
Our data

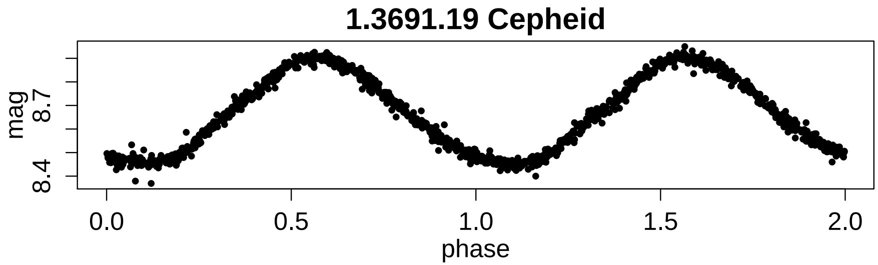
Our “training” data is a tiny subset of the MACHO light curve catalog.

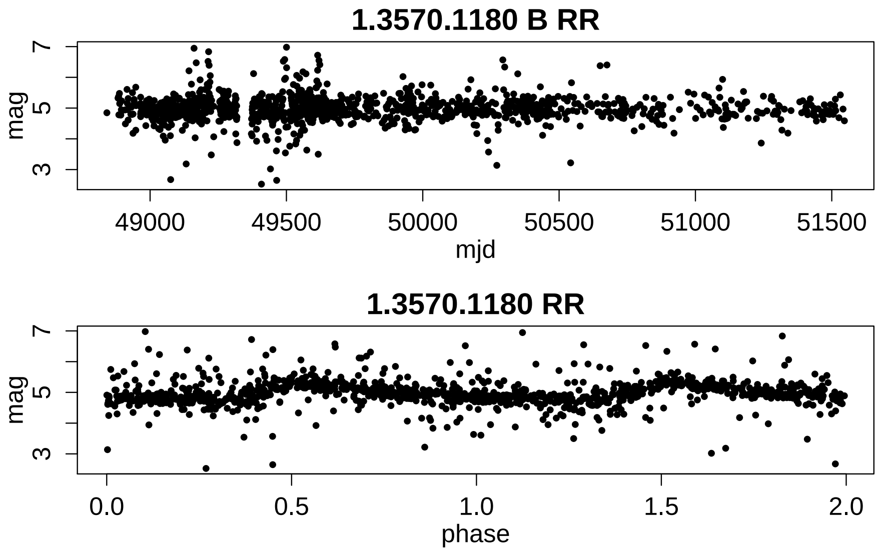
- 5652 number of curves
- 500-2000 observations per curve

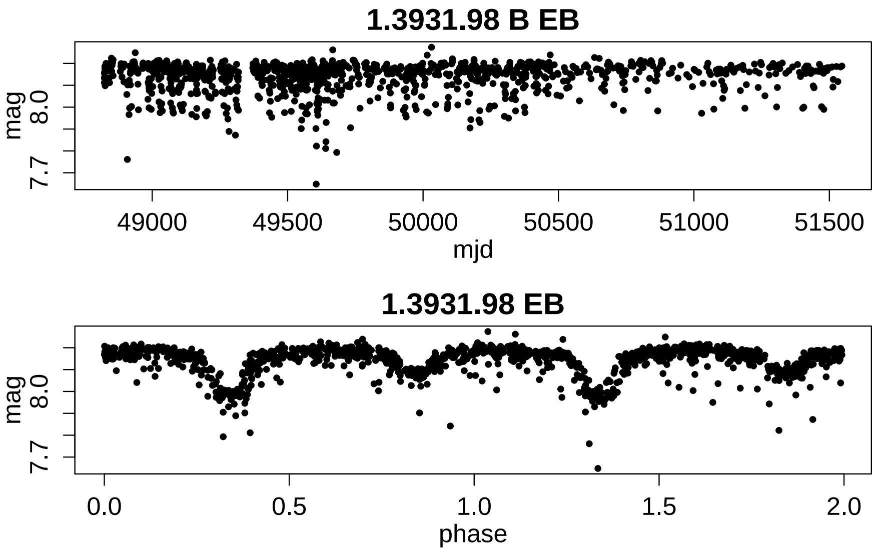
Types of variable sources in our data fall into three major categories:

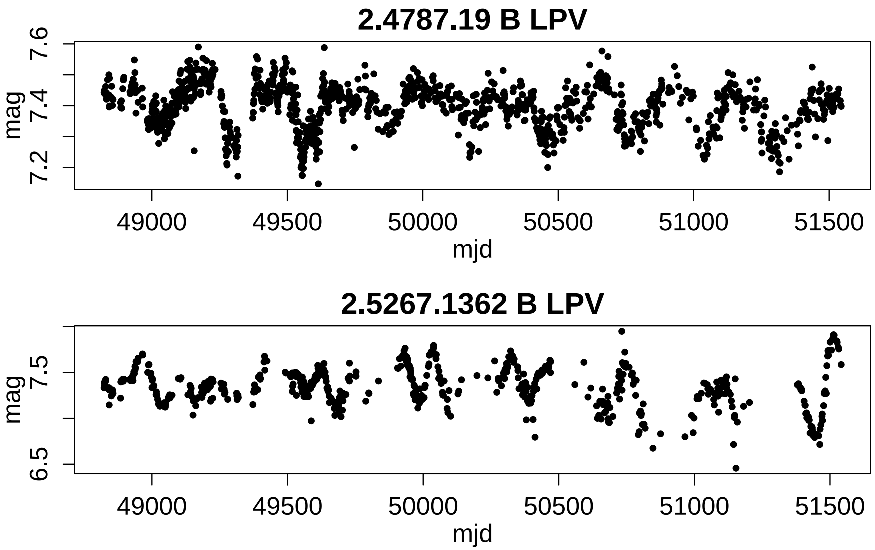
- Periodic sources: cepheids (short-period variable stars), eclipsing binary systems (EB), RR Lyrae, and long period variables (LPV).
- Non-periodic, stochastic sources: Be, Quasars.
- Event-based: Supernovae, microlensing events.
- (There are also nonvariable sources, which make up the majority of our database).

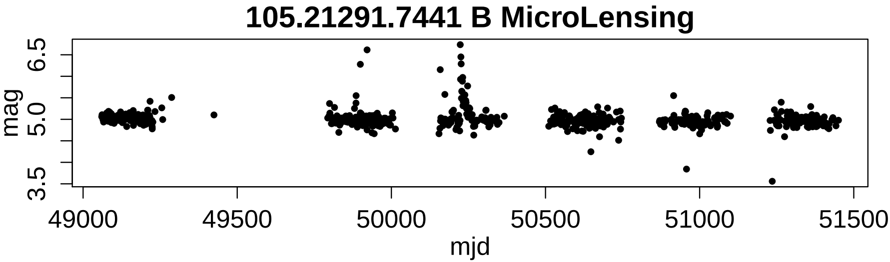

1.3691.19 B Cepheid
mag
8.4
8.7
49000
49500
50000
50500
51000
51500
mjd

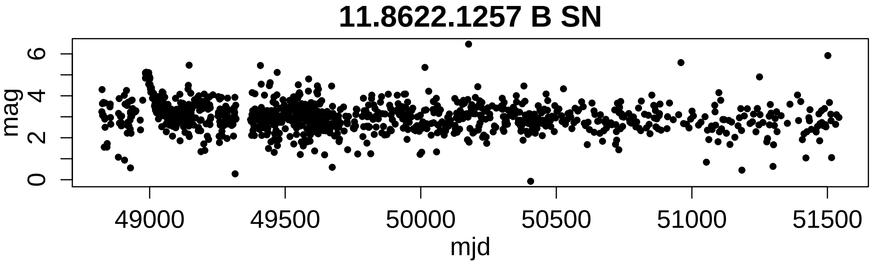

1.3691.19 Cepheid
mag
8.4
8.7
0.0
0.5
1.0
1.5
2.0
phase

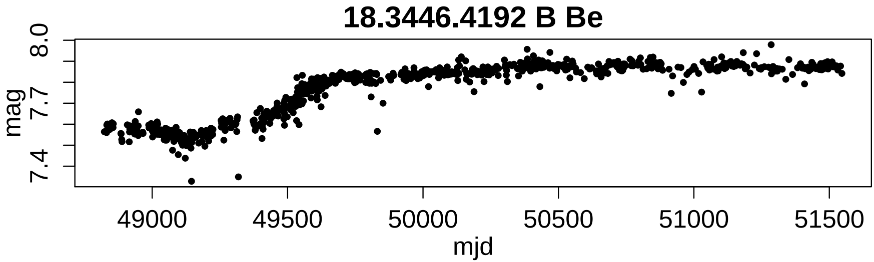

1.3570.1180 B RR
mag
7
5
3
49000
49500
50000
50500
51000
51500
mjd
1.3570.1180 RR
mag
7
5
3
0.0
0.5
1.0
1.5
2.0
phase

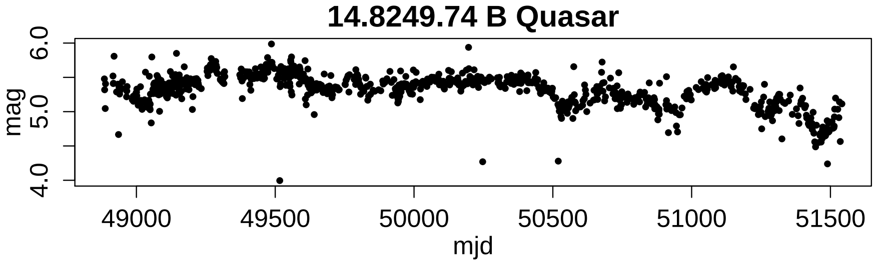

1.3931.98 B EB
mag
8.0
7.7
49000
49500
50000
50500
51000
51500
mjd
1.3931.98 EB
mag
8.0
7.7
0.0
0.5
1.0
1.5
2.0
phase

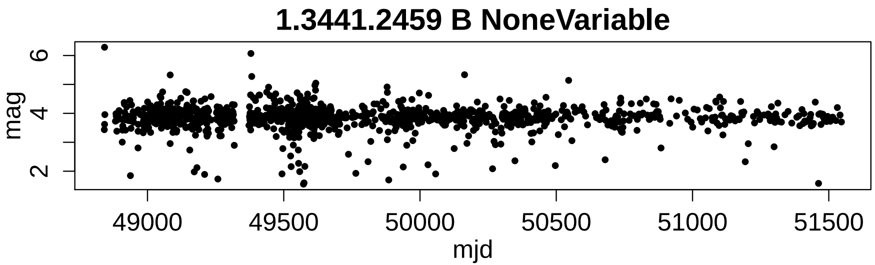

2.4787.19 B LPV
mag
7.2
7.4
7.6
49000
49500
50000
50500
51000
51500
mjd
2.5267.1362 B LPV
mag
6.5
7.5
49000
49500
50000
50500
51000
51500
mjd


105.21291.7441 B MicroLensing
mag
6.5
5.0
3.5
49000
49500
50000
50500
51000
51500
mjd

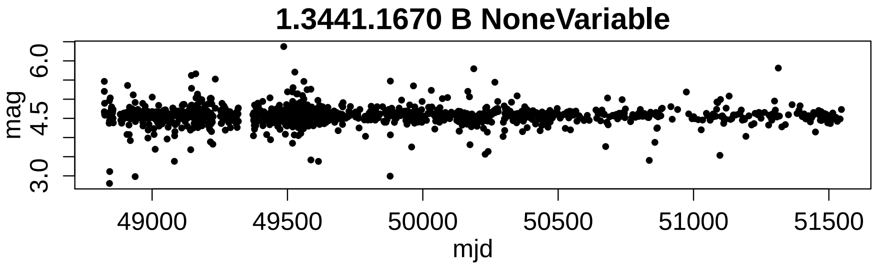

11.8622.1257 B SN
mag
6
4
2
0
49000
49500
50000
50500
51000
51500
mjd


18.3446.4192 B Be
mag
8.0
7.7
7.4
49000
49500
50000
50500
51000
51500
mjd


14.8249.74 B Quasar
mag
6.0
5.0
4.0
49000
49500
50000
50500
51000
51500
mjd



1.3441.2459 B NoneVariable
mag
2
4
6
49000
49500
50000
50500
51000
51500
mjd



1.3441.1670 B NoneVariable
mag
3.0
4.5
6.0
49000
49500
50000
50500
51000
51500
mjd

Model Blueprint

We seek a statistical procedure that simultaneously satisfies four goals

1. Classify an observed light curve, both for large and small numbers of observations.
2. Predict future observations of a light curve.
3. Use (1) and (2) to predict the time at which a future observation will be most informative.
4. Decision framework for use of the telescope.

Parts (1)-(3) will be adressed here in the context of our data, which represents only a subset of the variable source population. The decision framework alluded to in (4) would be an extension of the forthcoming results to reflect more specific scientific goals.

Classification

Classifying variable sources is a very active research topic in astronomy and astrostatistics. We used Random Forest classifiers because:

- Provide “soft” classification, which is necessary for our larger inferential procedure.
- Common choice in light curve classification literature, using features similar to what are extractable from our data.
- Relatively quick to train and use for prediction.

Classification

Features used for classification:

- Periodic features from generalized Lomb-Scargle periodogram:
  - Period, amplitude.
  - Variance reduction and goodness of fit.
  - Repeated at first harmonic.
- First four sample moments.
- Percentage of points beyond 1 SD of mean.
- Ratios of quantiles.

Classification

For those unfamiliar with a Random Forest classifier:

- "Forest" of classification trees, each tree trained on random subset of total training data.
- Randomly sample a small number of input variables to make decisions at each node of each tree.
- Repeat to grow a forest of trees.
- New inputs are passed through each tree, and their votes are averaged to obtain predicted class probabilities.
- Unbiased estimate of global error rates obtained by passing units through trees they didn't help build.

Classification

RF classifier confusion matrix, trained on 5 observations per light curve:

	ceph	rr	eb	lpv	be	qu	sn	mic	nv	class.error
ceph	50	1	19	8	0	0	0	0	0	0.36
rr	1	227	20	1	1	1	0	9	28	0.21
eb	10	50	90	32	2	0	0	6	3	0.53
lpv	3	11	32	283	17	2	0	8	5	0.22
be	0	1	8	84	17	3	0	8	6	0.87
qu	0	3	2	6	2	6	0	20	19	0.90
sn	0	1	0	0	0	1	0	1	5	1.00
mic	0	16	6	10	6	4	0	271	87	0.32
nv	0	12	3	12	4	1	0	78	290	0.28

Classification

RF classifier confusion matrix, trained on 50 observations per light curve:

	ceph	rr	eb	lpv	be	qu	sn	mic	nv	class.error
ceph	75	0	3	0	0	0	0	0	0	0.04
rr	0	261	14	0	0	0	0	6	7	0.09
eb	2	10	139	11	5	1	0	7	18	0.28
lpv	0	0	2	337	19	0	0	3	0	0.07
be	0	2	8	28	74	3	0	11	1	0.42
qu	0	4	3	5	4	10	0	24	8	0.83
sn	0	0	0	1	0	1	0	5	1	1.00
mic	0	9	2	9	12	2	0	343	23	0.14
nv	0	6	13	0	5	0	0	17	359	0.10

Prediction

We model the observed magnitudes as a latent Gaussian Process with additive, independent noise. Conditional on a source belonging to class c, for $i = 1, ..., n$, we observe magnitude y_i at time t_i, assuming:

- $y_i = f_i + \epsilon_i$
- $\epsilon_i \overset{iid}{\sim} N(0, V_i)$ with V_i known.
- $\mathbf{f} \sim N(\mu\mathbf{1}, K_c(\mathbf{t}, \mathbf{t}; \phi))$ where K_c is a covariance function corresponding to class c, parameterized by ϕ.

Why model the latent source intensity as a Gaussian Process?

- Smoothness.
- Can incorporate physical assumptions such as stationarity and periodicity.
- Computationally fast when using small samples and assuming additive Gaussian noise.

Prediction

We will use two covariance functions, one for classes with periodic sources and one for nonperiodic source classes.

$$\text{Squared exponential: } K_c(s,t;\phi) = \sigma^2 \exp(-\beta(t-s)^2)$$

$$\text{Periodic: } K_c(s,t;\phi) = \sigma^2 \exp\left(-\beta \sin\left(\frac{\pi(t-s)}{\tau}\right)^2\right)$$

- Both are isotropic (are functions only of $|t-s|$).
- σ^2 is the variance of the stationary distribution for the source intensity
- β is the (inverse) length-scale: larger values correspond to more variability in the source intensity per unit time; values closer to 0 correspond to smoother curves.

Prediction

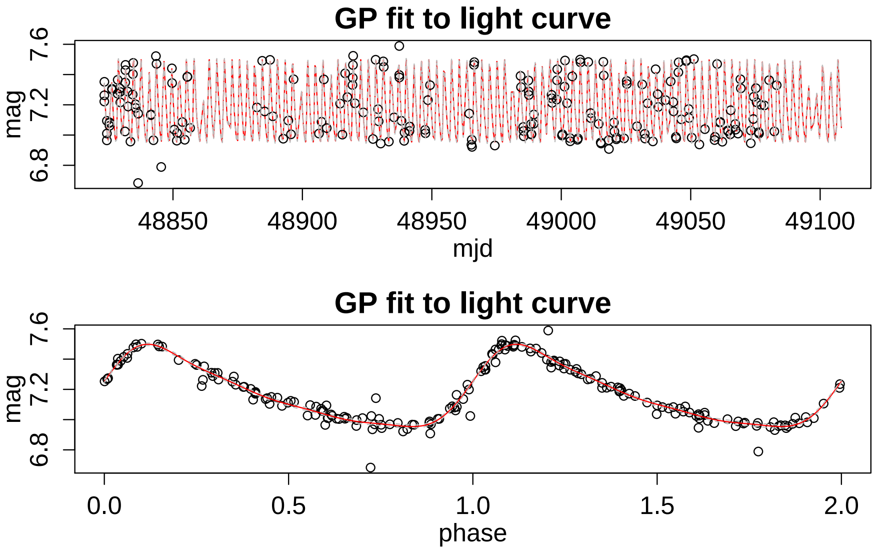
For a curve belonging to class c and the parameters μ and ϕ fixed, the predictive distribution for a future observation t^* is easily obtained:

$$\begin{pmatrix} \mathbf{y} \\ y^* \end{pmatrix} | C, \phi \sim N\left(\mu\mathbf{1}, \begin{pmatrix} K_c(\mathbf{t}, \mathbf{t}; \phi) + \mathbf{D_V} & K_c(t^*, \mathbf{t}; \phi) \\ K_c(\mathbf{t}, t^*; \phi) & \sigma^2 + V^* \end{pmatrix}\right)$$

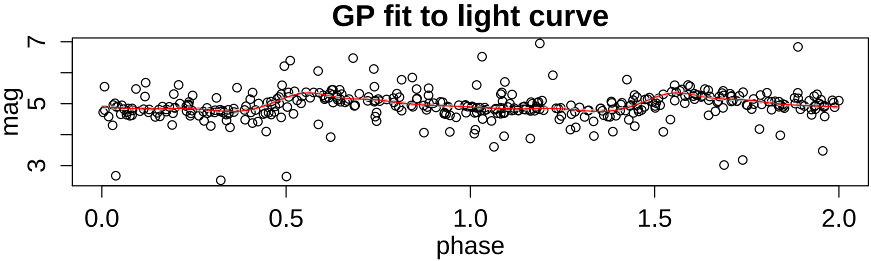
where $\mathbf{D_V} = \texttt{diag}(V_1, ..., V_n)$. V^* is unknown, but we may draw one from an inverse chi square or sample an existing observed V_i. Multivariate normal properties thus give

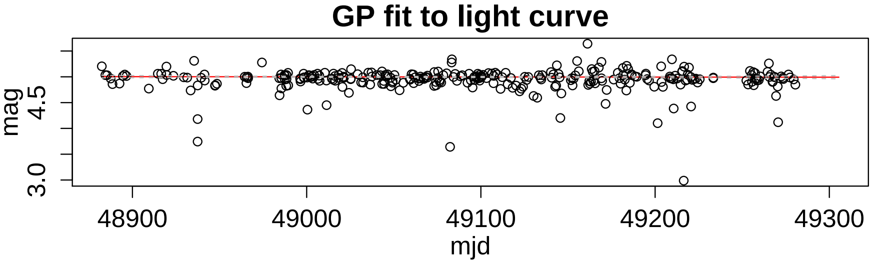
$$y^*|\mathbf{y}, V^*, C, \phi \sim N\left(\mu + K_{21}K_{11}^{-1}(\mathbf{y} - \mu\mathbf{1}), \sigma^2 + V^* - K_{21}K_{11}^{-1}K_{12}\right)$$

Prediction: GP fit for cepheids

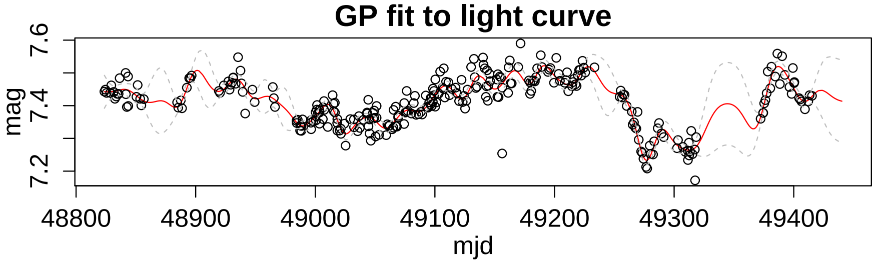


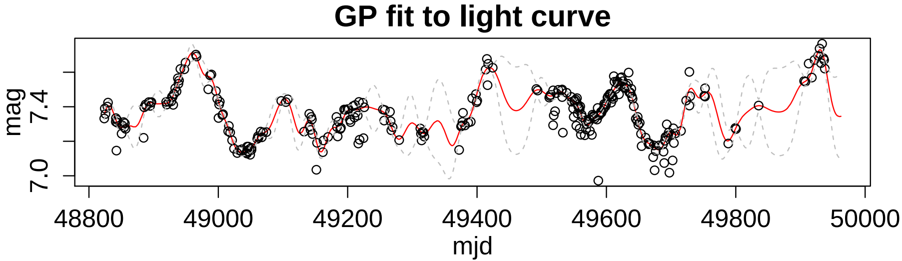
Prediction: GP fit for RR and none-variable



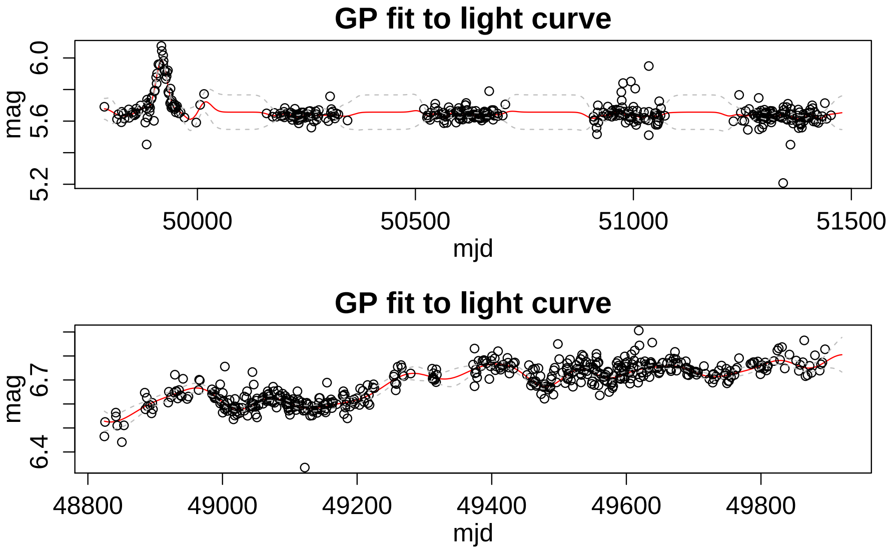


Prediction: GP fit for LPVs





Prediction: GP fit for Mic and Qu



Choosing future observations

Define the entropy for the multinomial distribution of class membership, conditional on the observed light curve:

$$H(C|\mathbf{y}) = -\sum_{c} P(C = c|\mathbf{y}) \log(P(C = c|\mathbf{y})) \tag{1}$$

For the purposes of classification, small entropies are desirable.

We define a related quantity, the **conditional entropy**, $H(C|y^*, \mathbf{y})(t^*)$, using (1) assuming we have a future observation y^*, and then averaging over the posterior predictive distribution $y^*|\mathbf{y}$:

$$H(C|y^*, \mathbf{y})(t^*) = \int_{-\infty}^{\infty} H(C|\mathbf{y}, y^*) p(y^*|\mathbf{y}) dy^* \tag{2}$$

This posterior predictive distribution $p(y^*|\mathbf{y})$ averages over unknown parameters of the Gaussian Process model of the source intensity as well as the unknown class memberships.

Choosing future observations

Why consider conditional entropy $H(C|y^*, \mathbf{y})(t^*)$?

- Function only of t^*; represent mean information gained for classificatoin by observing next at time t^*.
- How are future observations useful to use if they are imputed from the present?
- Equivalent to considering mutual information for future observation y^* and class identity variable C, conditional on observed data.

Summary of inferential procedure

So in order to classify light curves as quickly as possible, we (after having observed a handful of points initially) we:

1. Obtain class probabilities conditional on observed data using RF classifier, $P(C|\mathbf{y})$.
2. Obtain posterior distributions of GP parameters μ, ϕ for each class (with nonzero probability).
3. Pick candidate t^* from a reasonable range of possible values given material constraints.
4. For this t^*, use (1)-(2) to sample from the posterior predictive distrubtion $p(y^*|\mathbf{y})$.
5. Using these samples, compute the conditional entropy $H(C|y^*, \mathbf{y})$.
6. Iterate steps (3)-(5) through your candidate set for t^*.
7. Set $t_{n+1} = \text{argmin}_{t^*} H(C|y^*, \mathbf{y})$ and make observation.
8. Repeat.

Choice of prior

Drawing from the posterior predictive distribution involves samplng from $p(\mu, \phi|\mathbf{y}, C = c)$ for all classes c. A priori, we assume

$$\begin{pmatrix} \mu \\ \log(\phi) \end{pmatrix} |C \sim N\left(\begin{pmatrix} \mu_{0,c} \\ \tilde{\phi}_{0,c} \end{pmatrix}, \Sigma_{0,c}\right)$$

($\tilde{\phi}$ represents $\log(\phi)$). For each class, we set $\mu_{0,c}, \tilde{\phi}_{0,c}, \Sigma_{0,c}$ by

- Choosing a random subset of the light curves from class c and finding the MLEs for μ and $\tilde{\phi}$ for each using all observations.
- Setting $\mu_{0,c}, \tilde{\phi}_{0,c}, \Sigma_{0,c}$ to the sample moments.
- Should give similar results as maximal marginal likelihood but much easier to implement.

Sampling from posterior

Sampling the posterior $p(\mu, \phi | \mathbf{y}, C = c)$ requires the following considerations:

- Needs to be efficient; every evaluation of the likelihood (and its gradient) requires matrix inversion.
- Should require no "hand" tuning, as we want it to run sequentially across sets of candidate observations over time.
- Handles multimodality; this is very common especially for the periodic kernel.

Metropolis-Hastings algorithm:

- Locate posterior modes and calculate first two derivatives.
- Using heights and curvature at modes, fit a multivariate t mixture approximation for the posterior.
- Generate independent Metropolis-Hastings proposals from this approximation to the posterior.

Rules of probability and information theory

Combining fully parameterized Bayesian model for observations with nonparametric feature-based classifier has several consequences:

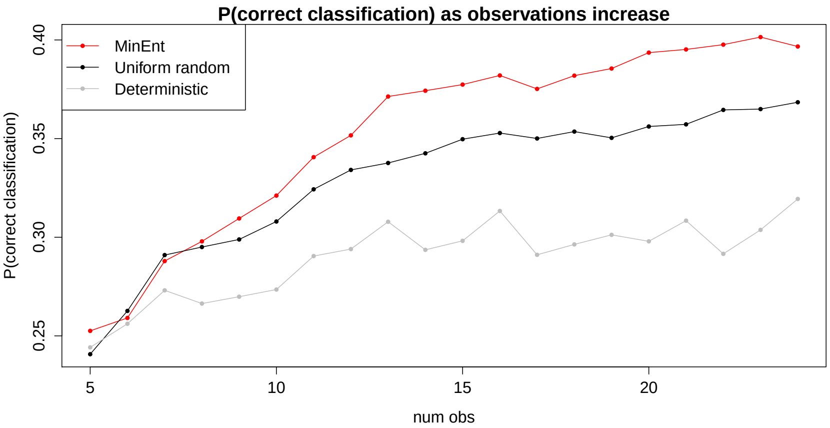
- Does the class-conditional distribution of features for each curve type depend on the observation schedule? This may bias $P(C|\mathbf{y})$.
- What is the joint probability for $p(y^*, C|\mathbf{y})$? Two unequal representations depending on what is conditioned on:
  - $p(y^*|\mathbf{y}, C = c)P(C = c|\mathbf{y}) \neq P(C = c|\mathbf{Y}, y^*)p(y^*|\mathbf{y})$.
- Information additivity does not hold.
  - Theoretically $H(C|\mathbf{y}, y^*) \leq H(C|\mathbf{y})$.
  - This will not always hold with our model.
- Could this invite disaster?

Results

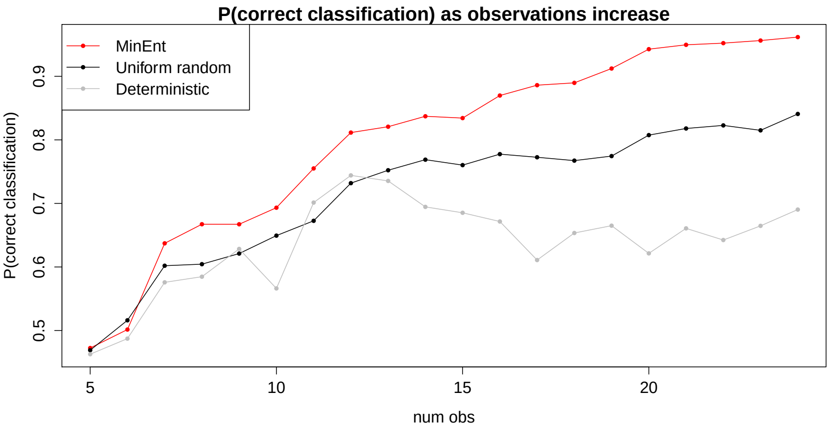
Our results are based on simulated light curves.

- 9 “fake” curves for each class.
- For each curve, model for providing noise variance for any given t.
- MinEnt observational design compared to deterministic observation schedule and random observation schedule.
- Metric of comparison is probability of correct classification vs number of observed points.

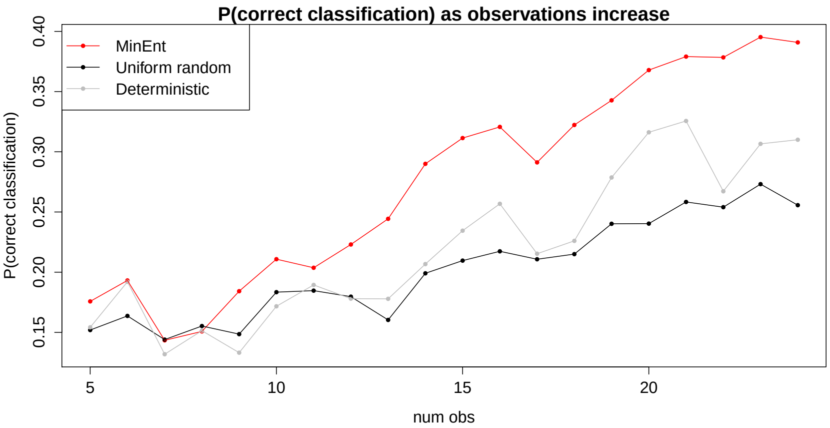
Correct classification probability (all types)



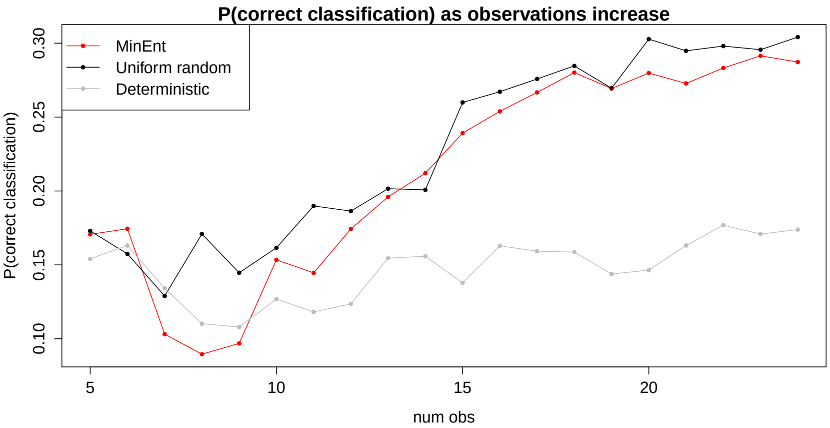
Correct classification probability (Cepheids)



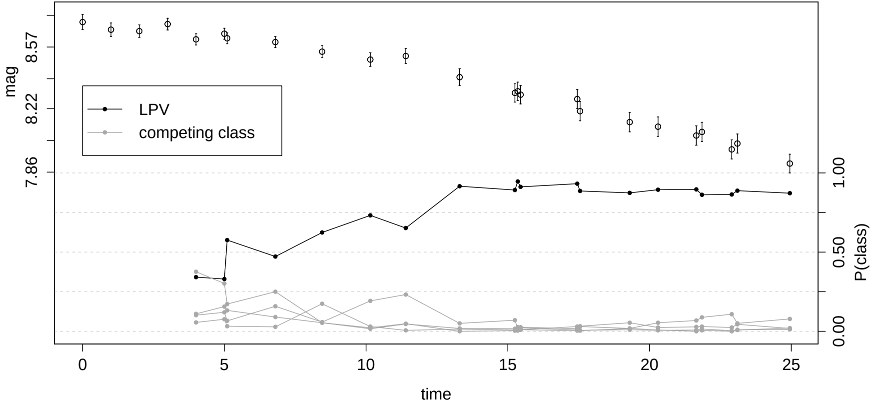
Correct classification probability (Be)



Correct classification probability (Eclipsing Binaries)



Example: observations on a LPV



Summary of results

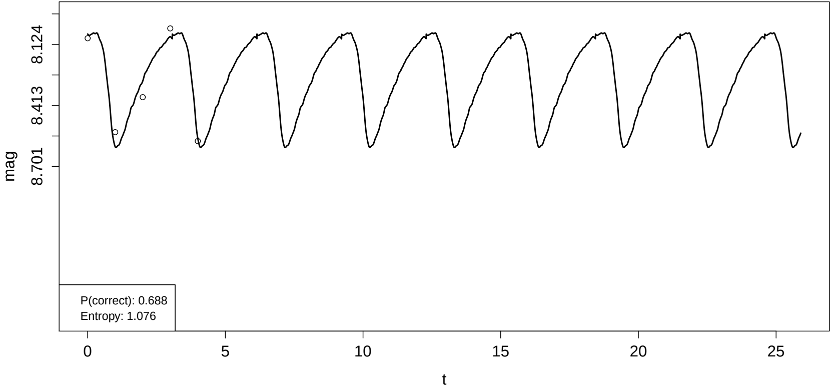
The MinEnt observational selection scheme presented here seems to be an improvement over arbitrary random or deterministic observation schedules.

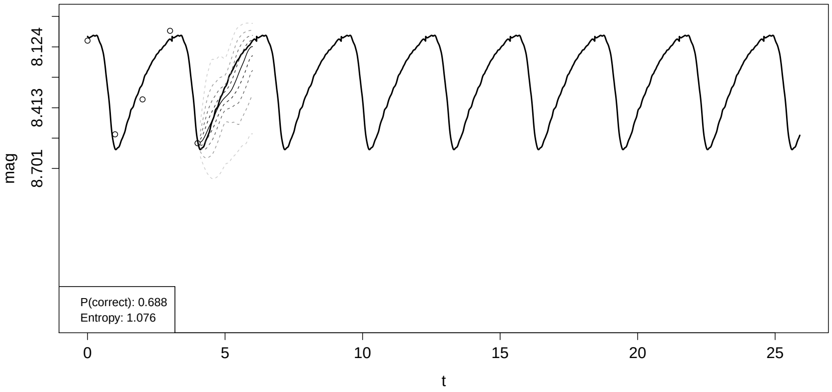
- True for measuring probability of correct classification over time (for most classes), as well as reduction in entropy over time.
- Strength of results hugely dependent on efficacy of classifier.
- We don't see improvements for classes whose features develop over longer time scale than what we use here.
- Results could also be strengthed by specificying more specific scientific goals/constraints (cost of time, different losses for different misclassifications).

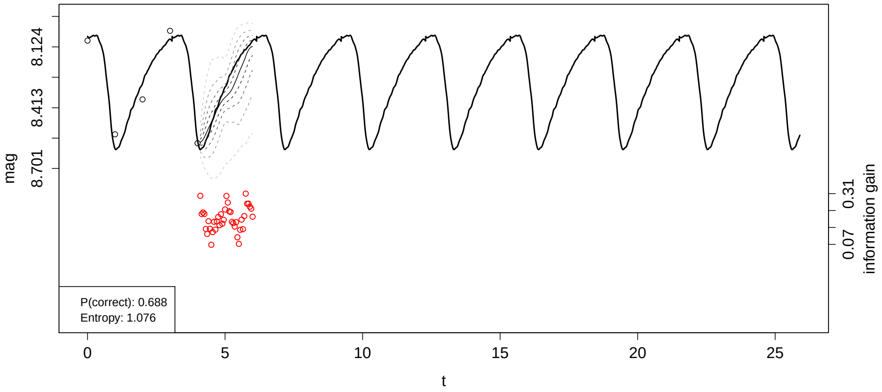
Caveats and future improvements

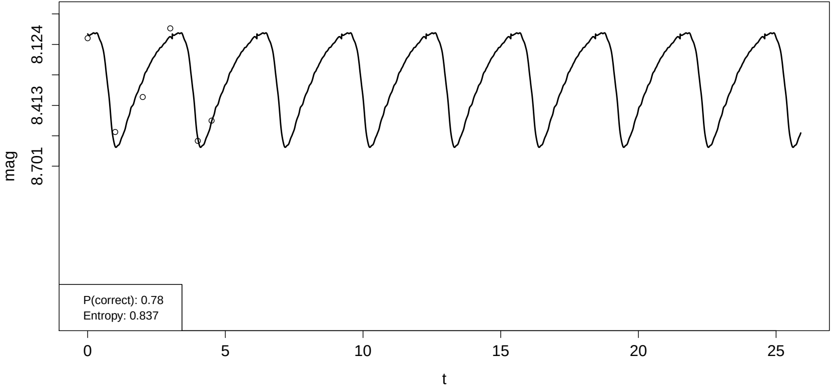
The following are ways in which the model could be improved:

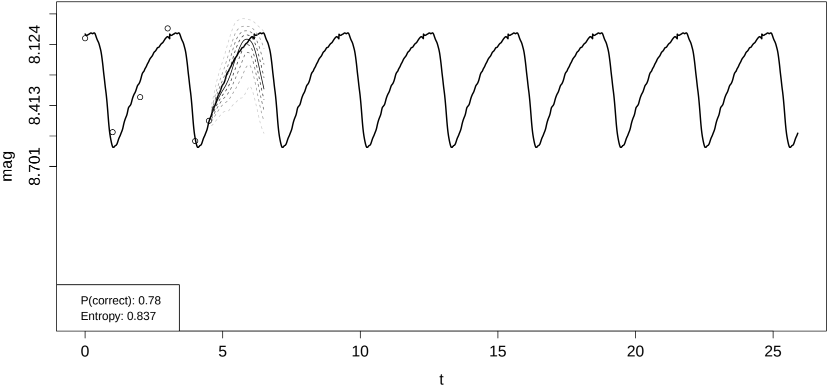
- Different modeling for additive noise (not actually independent of source intensity).
- Sequentially updating RF classifier, population distributions for μ, ϕ.
- Incorporating event detection procedures in features used for classification, and also in prediction.
- Incorporating observations from different spectra.
- Scalability: will this work over longer candidate observation windows, and for a longer number of iterations?
- Can we detect a new class?

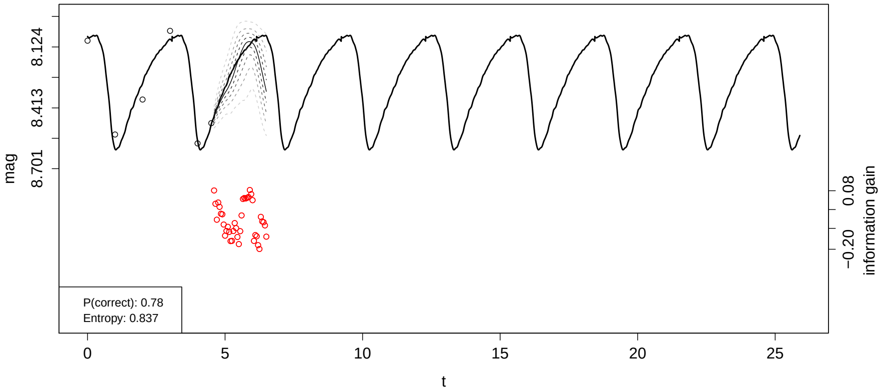

8.124
8.413
8.701
mag
P(correct): 0.688
Entropy: 1.076
0
5
10
15
20
25
t


8.124
8.413
8.701
mag
P(correct): 0.688
Entropy: 1.076
0
5
10
15
20
25
t

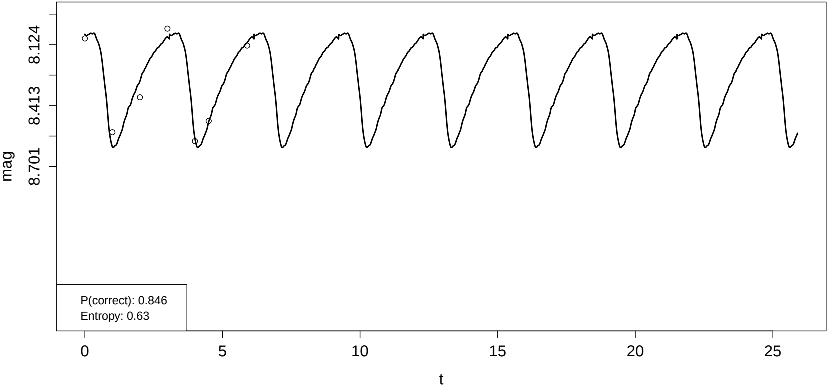

8.124
8.413
8.701
mag
0.31
0.07
information gain
P(correct): 0.688
Entropy: 1.076
0
5
10
15
20
25
t

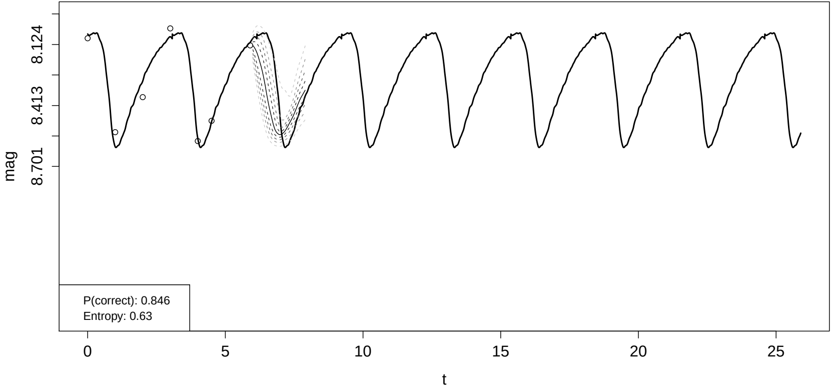

mag
8.124
8.413
8.701
P(correct): 0.78
Entropy: 0.837
0
5
10
15
20
25
t

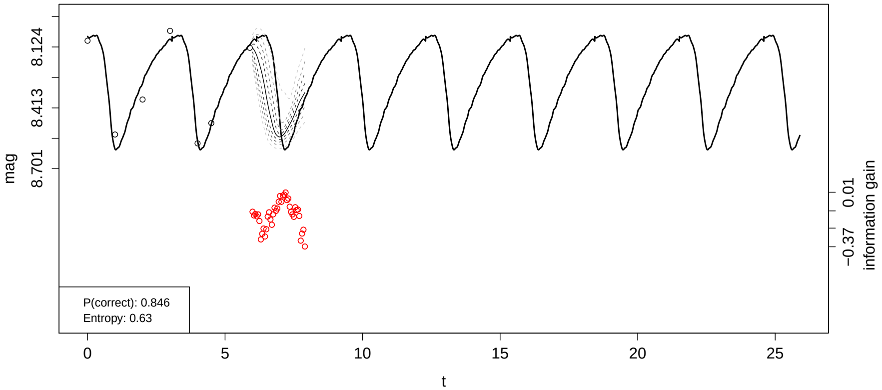

8.124
8.413
8.701
mag
P(correct): 0.78
Entropy: 0.837
0
5
10
15
20
25
t

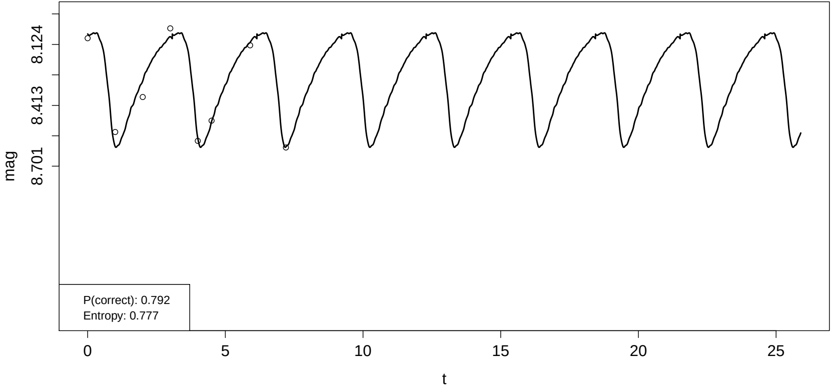


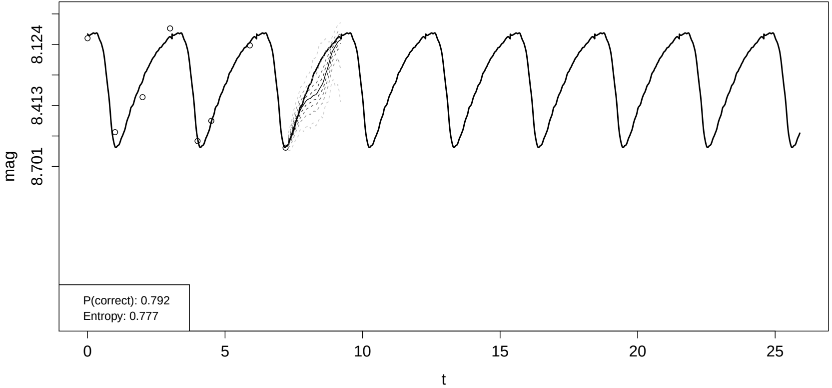
mag
8.124
8.413
8.701
0.08
−0.20
information gain
P(correct): 0.78
Entropy: 0.837
0
5
10
15
20
25
t

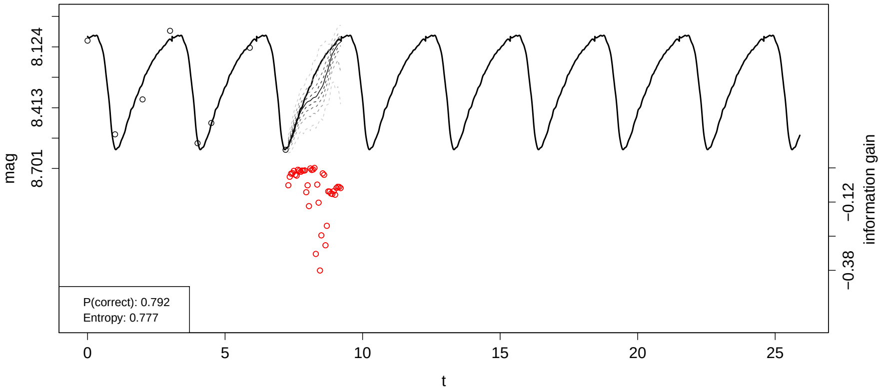

P(correct): 0.846
Entropy: 0.63
mag
8.124
8.413
8.701
0
5
10
15
20
25
t

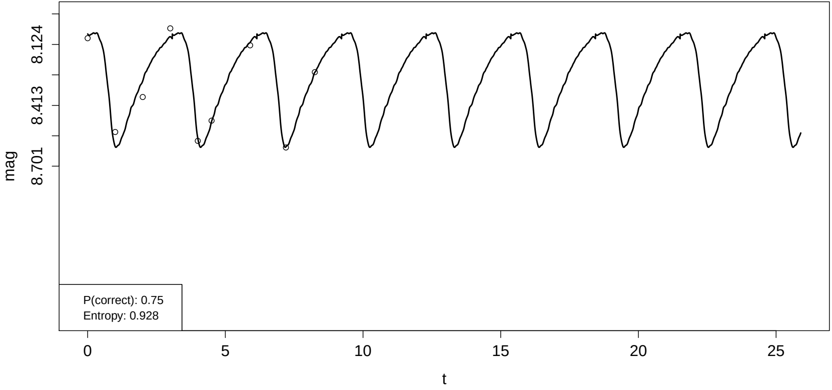

8.124
8.413
8.701
mag
P(correct): 0.846
Entropy: 0.63
0
5
10
15
20
25
t


mag
8.124
8.413
8.701
0.01
−0.37
information gain
P(correct): 0.846
Entropy: 0.63
0
5
10
15
20
25
t

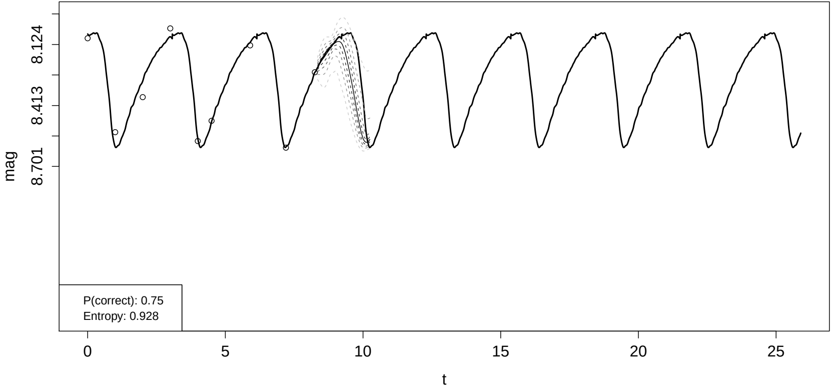

mag
8.124
8.413
8.701
P(correct): 0.792
Entropy: 0.777
0
5
10
15
20
25
t

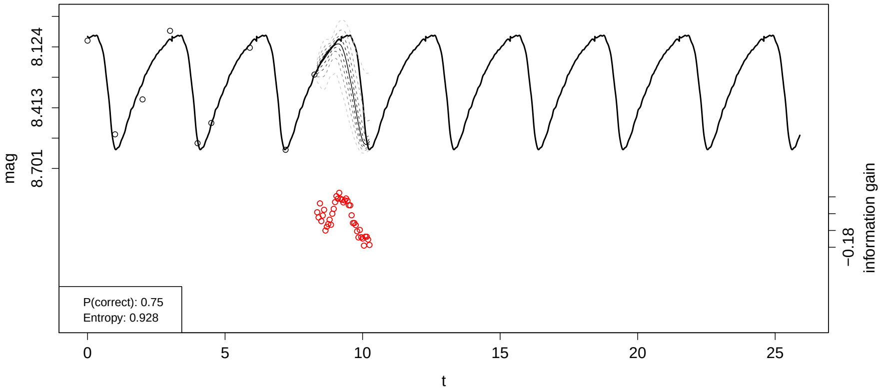

mag
8.124
8.413
8.701
t
0
5
10
15
20
25
P(correct): 0.792
Entropy: 0.777

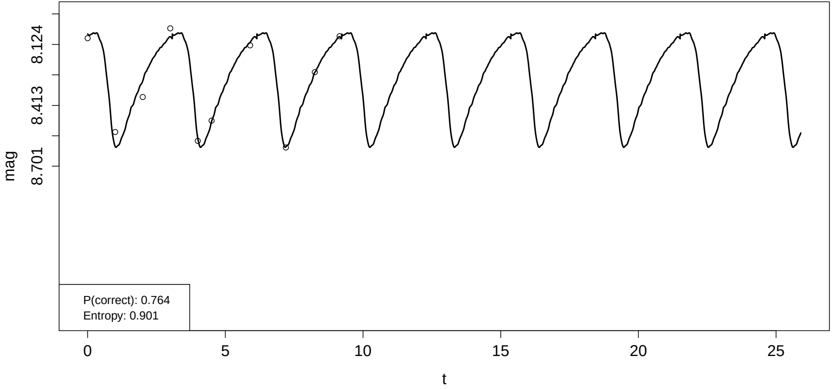

mag
8.124
8.413
8.701
information gain
−0.12
−0.38
P(correct): 0.792
Entropy: 0.777
0
5
10
15
20
25
t

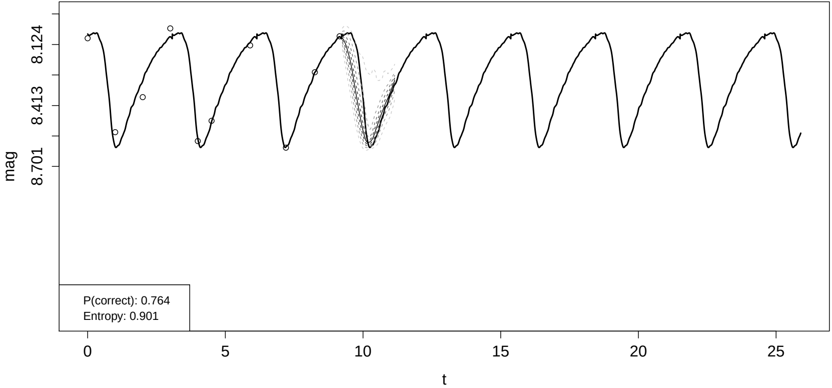


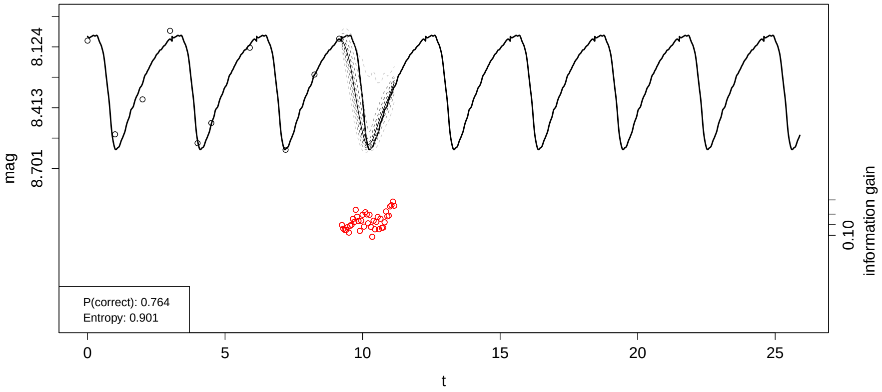
mag
8.124
8.413
8.701
P(correct): 0.75
Entropy: 0.928
0
5
10
15
20
25
t

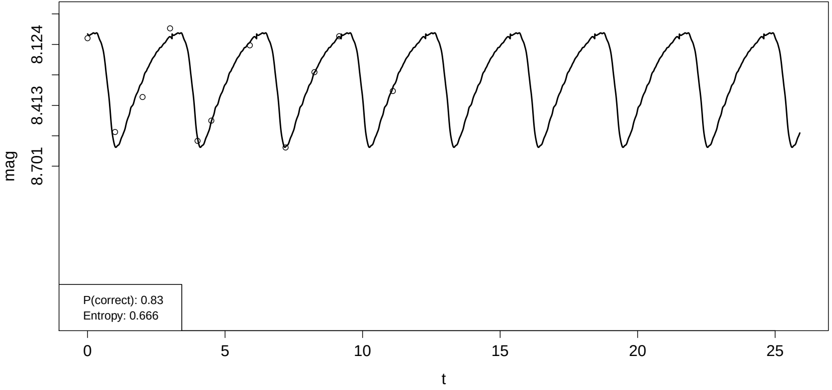

8.124
8.413
8.701
mag
0
5
10
15
20
25
t
P(correct): 0.75
Entropy: 0.928

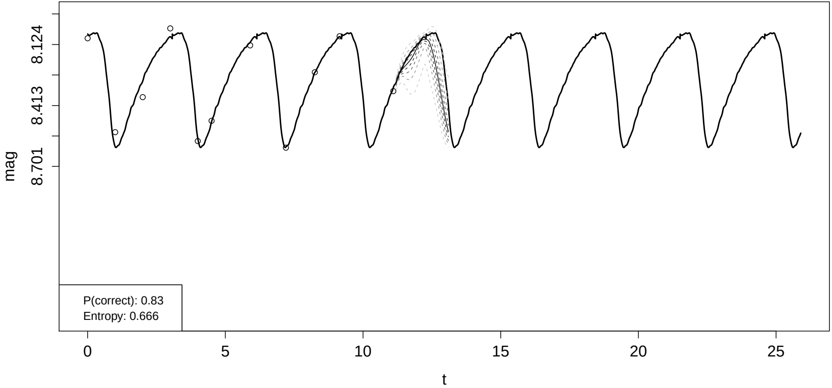

mag
8.124
8.413
8.701
information gain
−0.18
P(correct): 0.75
Entropy: 0.928
0
5
10
15
20
25
t


8.124
8.413
8.701
mag
P(correct): 0.764
Entropy: 0.901
0
5
10
15
20
25
t

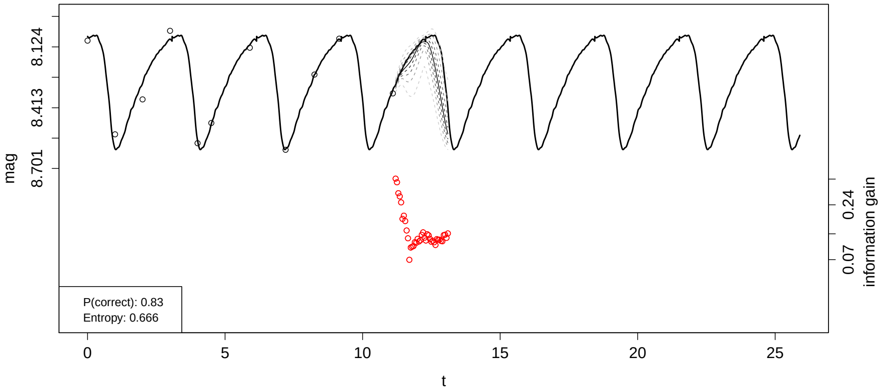

mag
8.124
8.413
8.701
t
0
5
10
15
20
25
P(correct): 0.764
Entropy: 0.901


mag
8.124
8.413
8.701
information gain
0.10
P(correct): 0.764
Entropy: 0.901
0
5
10
15
20
25
t

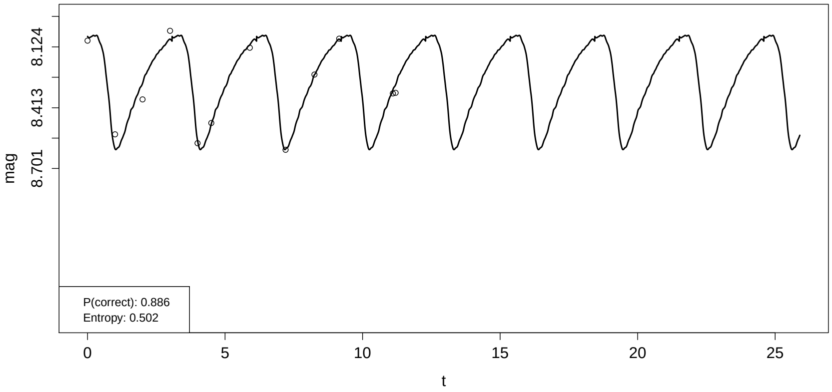

mag
8.124
8.413
8.701
P(correct): 0.83
Entropy: 0.666
0
5
10
15
20
25
t

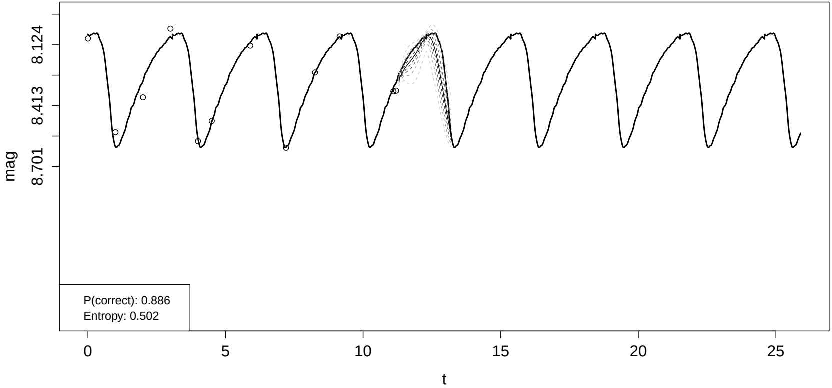


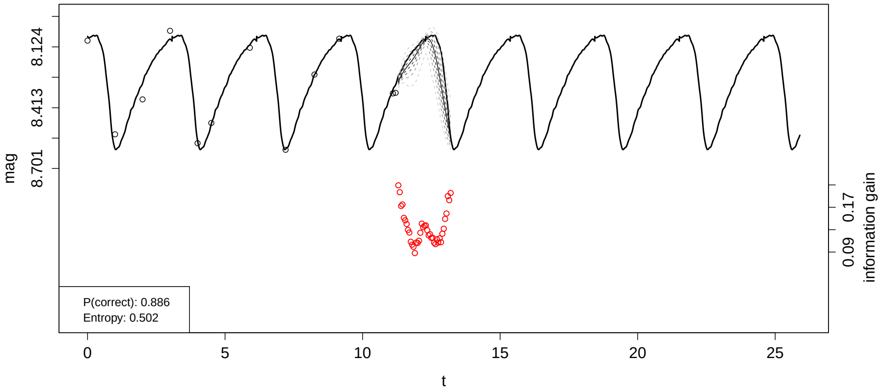
mag
8.124
8.413
8.701
P(correct): 0.83
Entropy: 0.666
0
5
10
15
20
25
t



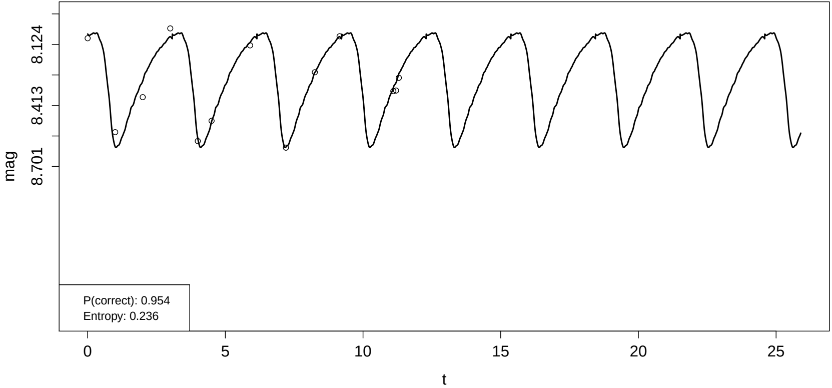
mag
8.124
8.413
8.701
information gain
0.24
0.07
P(correct): 0.83
Entropy: 0.666
0
5
10
15
20
25
t

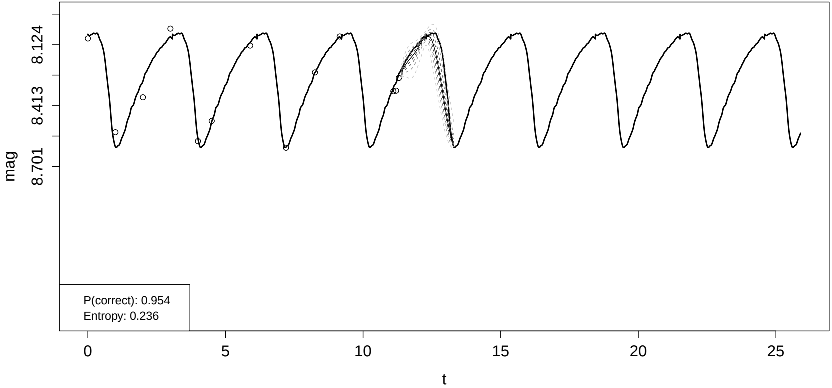

P(correct): 0.886
Entropy: 0.502
mag
8.124
8.413
8.701
0
5
10
15
20
25
t

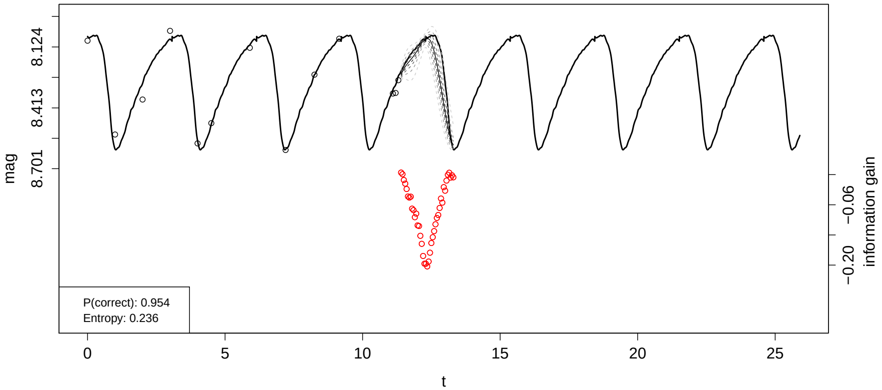

mag
8.124
8.413
8.701
t
0
5
10
15
20
25
P(correct): 0.886
Entropy: 0.502

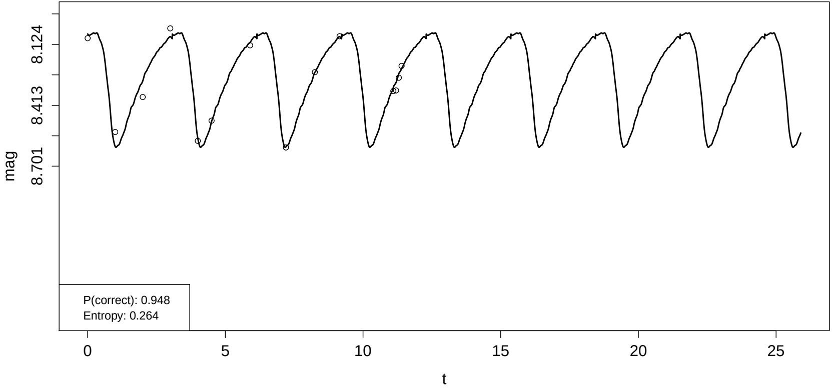


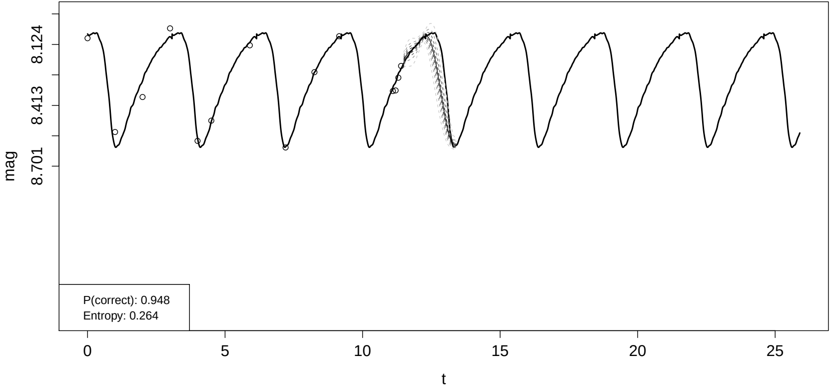
mag
8.124
8.413
8.701
information gain
0.09
0.17
P(correct): 0.886
Entropy: 0.502
0
5
10
15
20
25
t


mag
8.124
8.413
8.701
P(correct): 0.954
Entropy: 0.236
0
5
10
15
20
25
t

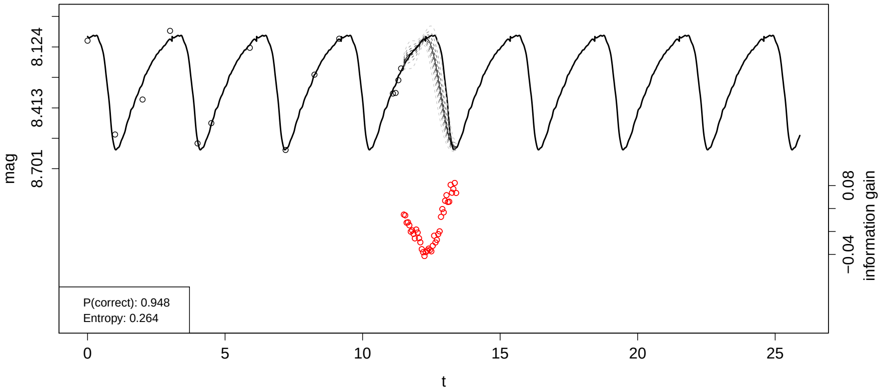

mag
8.124
8.413
8.701
0
5
10
15
20
25
t
P(correct): 0.954
Entropy: 0.236

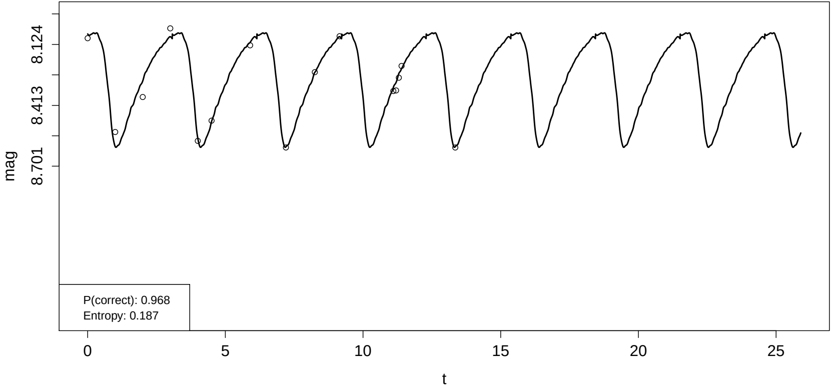

mag
8.124
8.413
8.701
information gain
−0.06
−0.20
t
0
5
10
15
20
25
P(correct): 0.954
Entropy: 0.236

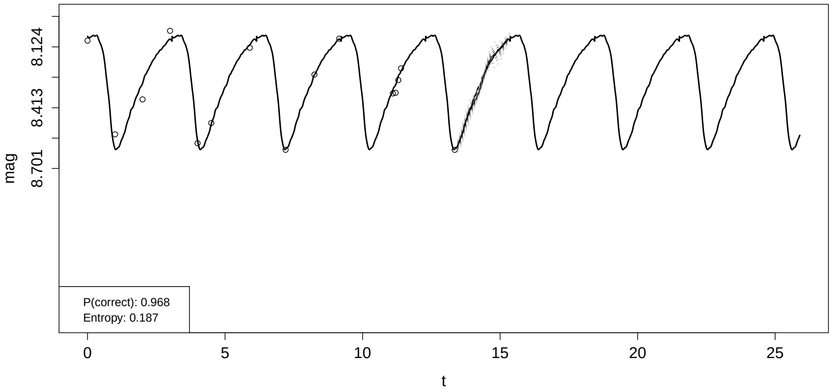

mag
8.124
8.413
8.701
0
5
10
15
20
25
t
P(correct): 0.948
Entropy: 0.264



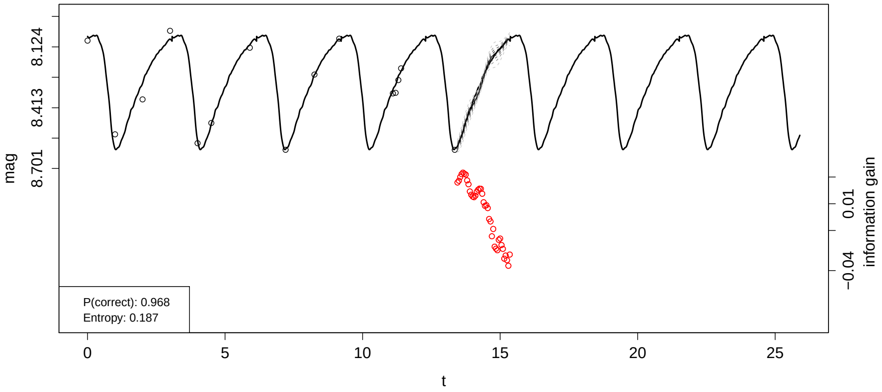
mag
8.124
8.413
8.701
P(correct): 0.948
Entropy: 0.264
0
5
10
15
20
25
t

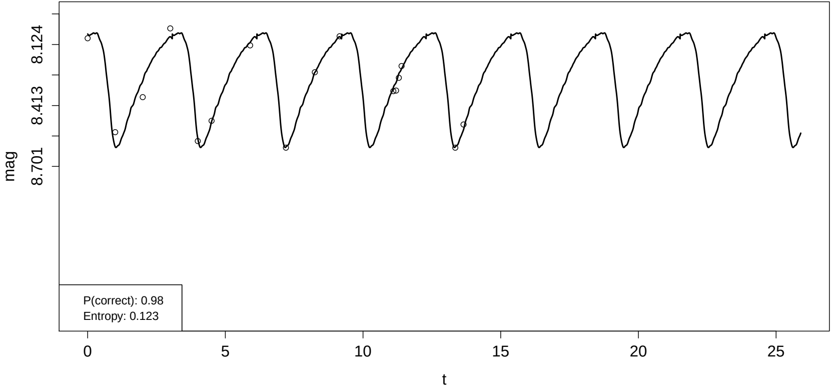

mag
8.124
8.413
8.701
0.08
−0.04
information gain
P(correct): 0.948
Entropy: 0.264
0
5
10
15
20
25
t

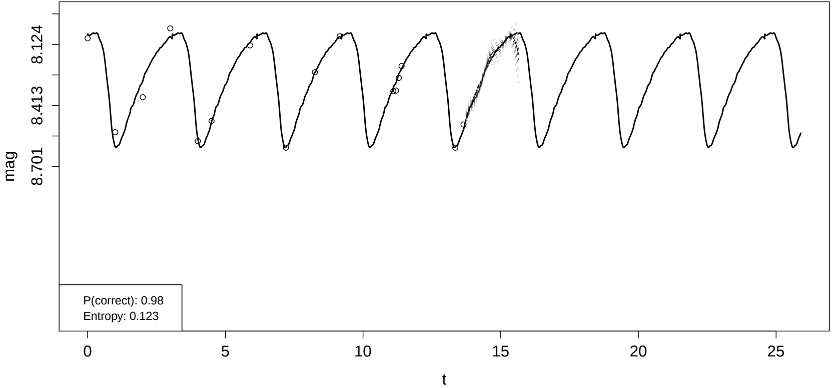

mag
8.124
8.413
8.701
t
0
5
10
15
20
25
P(correct): 0.968
Entropy: 0.187

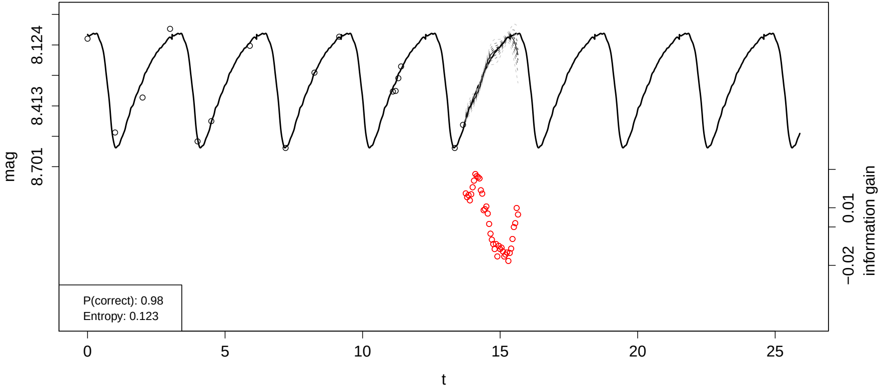


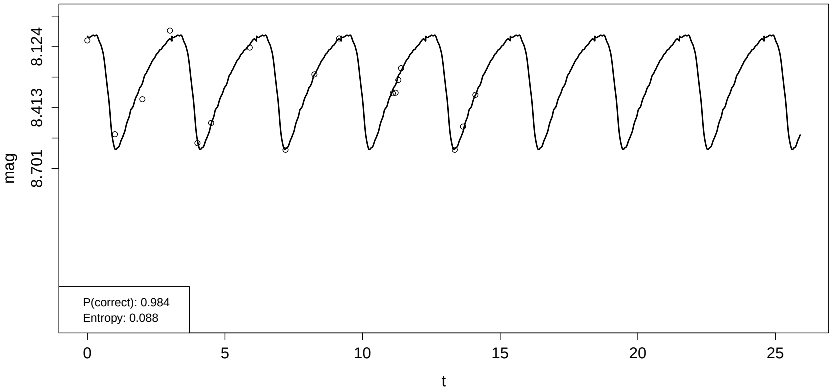
mag
8.124
8.413
8.701
t
0
5
10
15
20
25
P(correct): 0.968
Entropy: 0.187

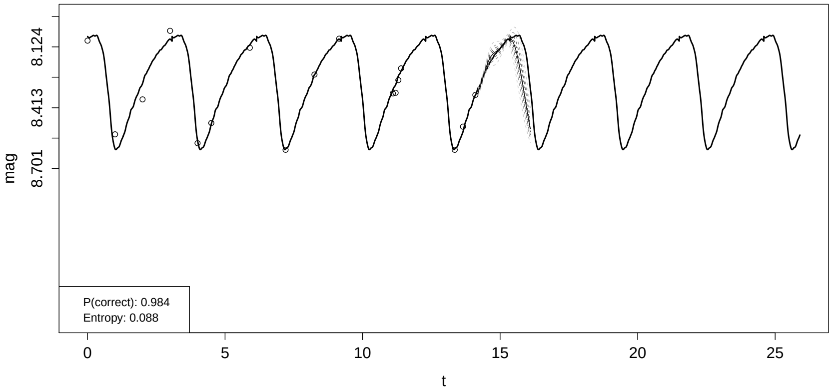

mag
8.124
8.413
8.701
information gain
0.01
−0.04
P(correct): 0.968
Entropy: 0.187
0
5
10
15
20
25
t

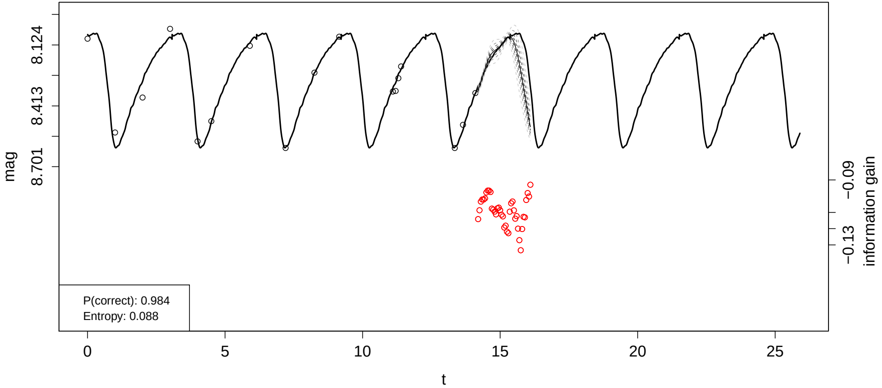

mag
8.124
8.413
8.701
0
5
10
15
20
25
t
P(correct): 0.98
Entropy: 0.123


mag
8.124
8.413
8.701
P(correct): 0.98
Entropy: 0.123
0
5
10
15
20
25
t

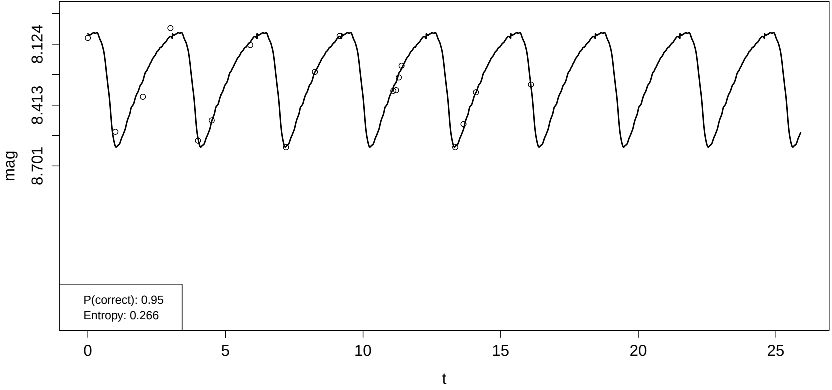

mag
8.124
8.413
8.701
information gain
0.01
−0.02
P(correct): 0.98
Entropy: 0.123
0
5
10
15
20
25
t

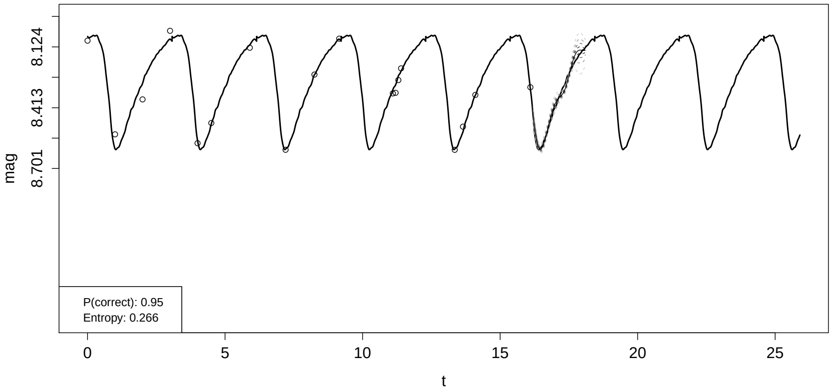

mag
8.124
8.413
8.701
P(correct): 0.984
Entropy: 0.088
0
5
10
15
20
25
t

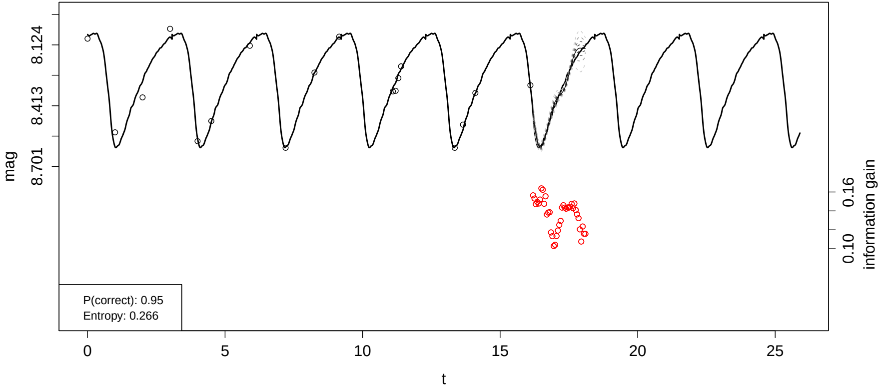

mag
8.124
8.413
8.701
0
5
10
15
20
25
t
P(correct): 0.984
Entropy: 0.088

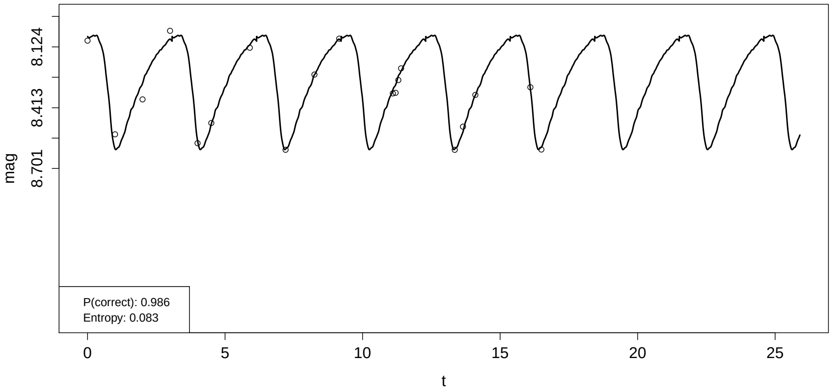


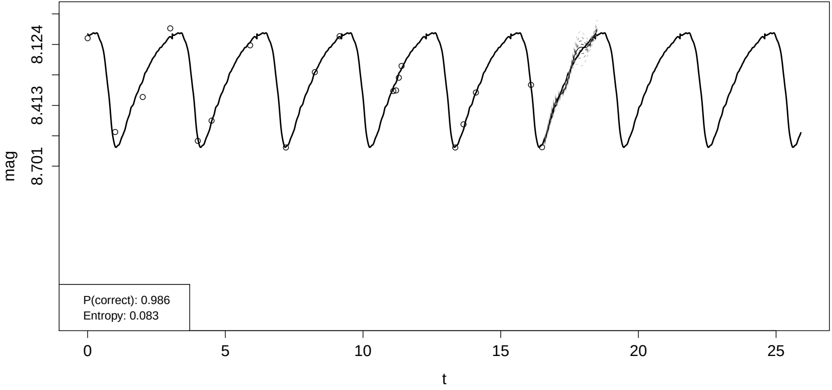
8.124
8.413
8.701
mag
0
5
10
15
20
25
t
−0.09
−0.13
information gain
P(correct): 0.984
Entropy: 0.088

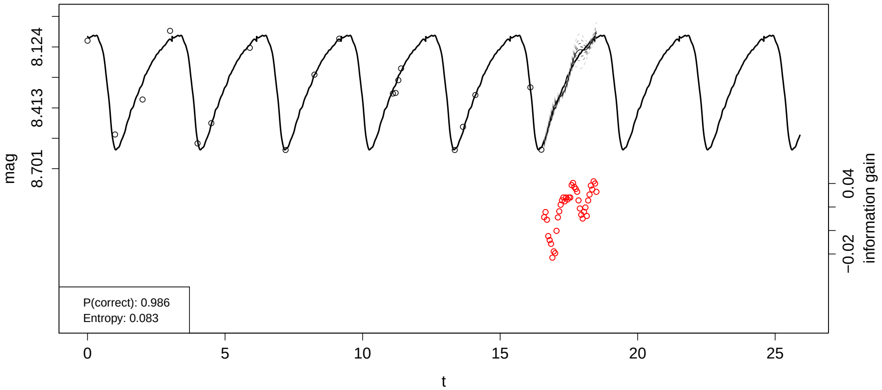

mag
8.124
8.413
8.701
0
5
10
15
20
25
t
P(correct): 0.95
Entropy: 0.266


mag
8.124
8.413
8.701
P(correct): 0.95
Entropy: 0.266
0
5
10
15
20
25
t

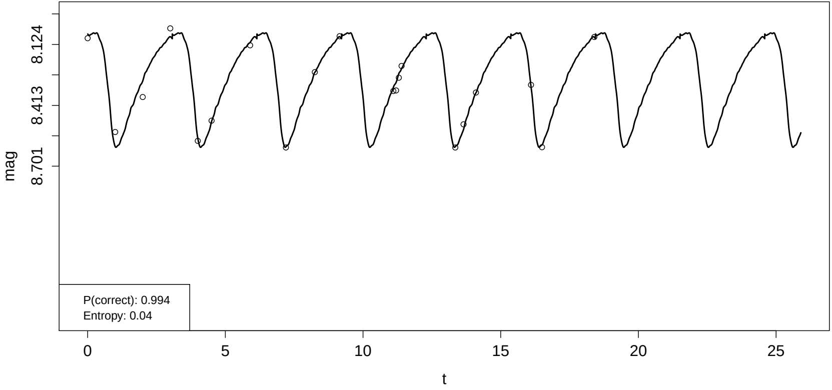

mag
8.124
8.413
8.701
information gain
0.16
0.10
P(correct): 0.95
Entropy: 0.266
0
5
10
15
20
25
t


mag
8.124
8.413
8.701
t
0
5
10
15
20
25
P(correct): 0.986
Entropy: 0.083

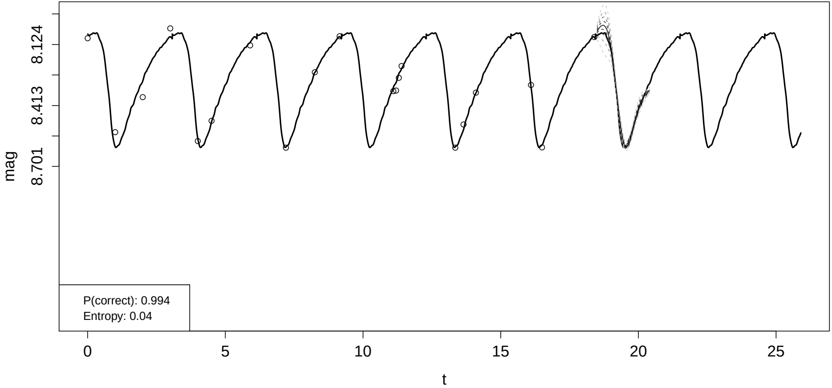

8.124
8.413
8.701
mag
0
5
10
15
20
25
t
P(correct): 0.986
Entropy: 0.083

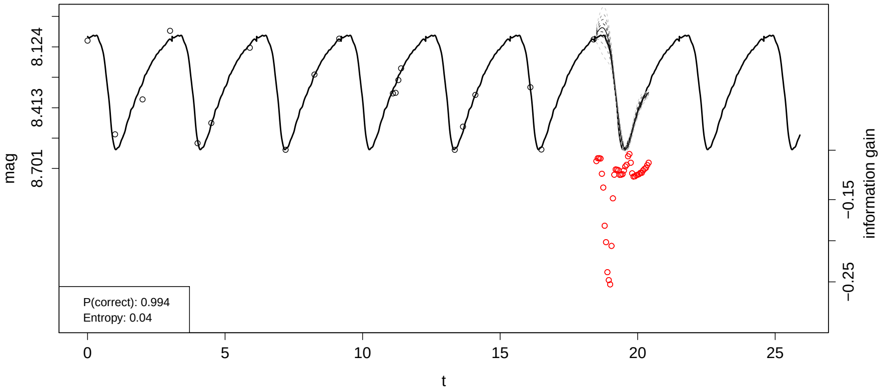


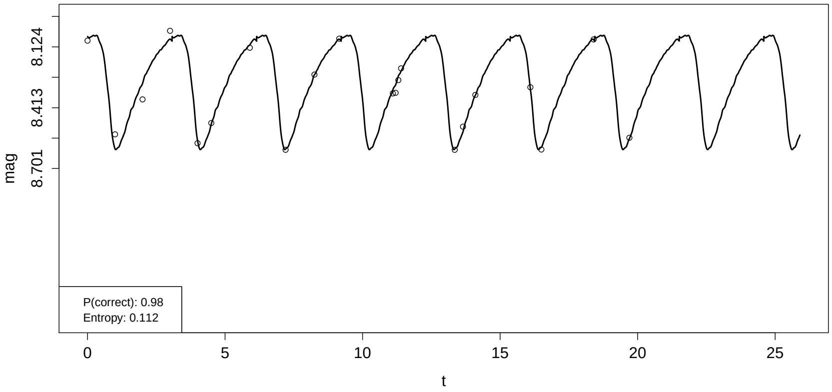
mag
8.124
8.413
8.701
information gain
0.04
−0.02
P(correct): 0.986
Entropy: 0.083
0
5
10
15
20
25
t

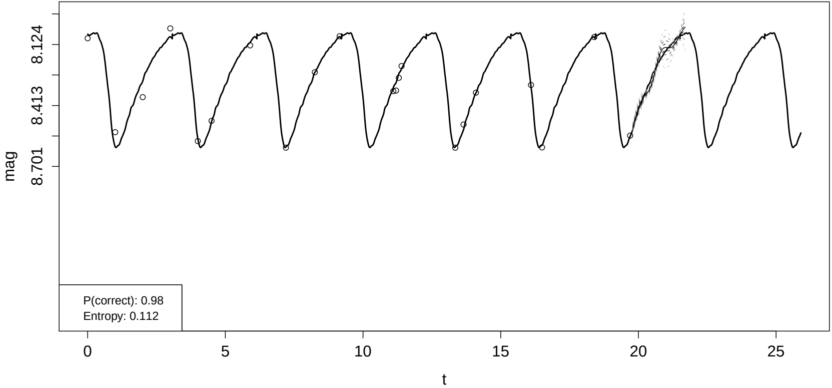


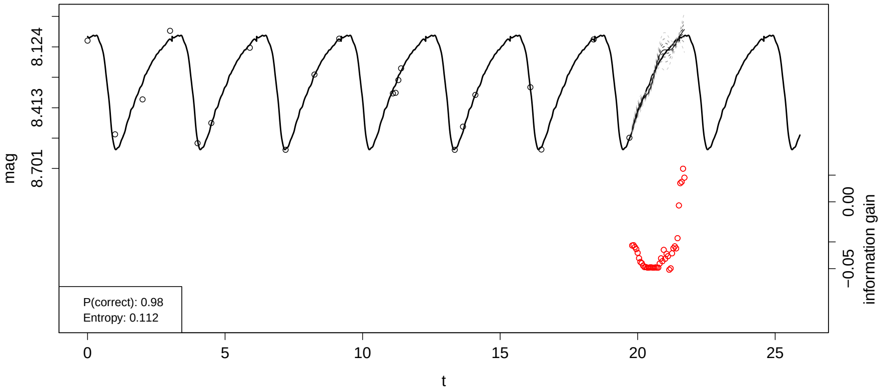
mag
8.124
8.413
8.701
t
0
5
10
15
20
25
P(correct): 0.994
Entropy: 0.04

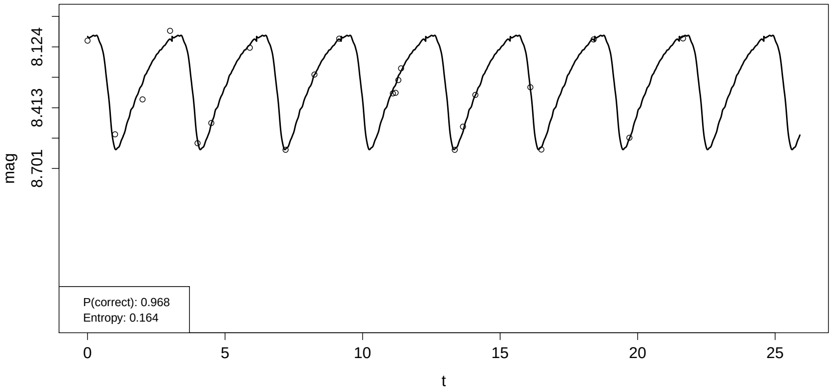

mag
8.124
8.413
8.701
t
0
5
10
15
20
25
P(correct): 0.994
Entropy: 0.04


8.124
8.413
8.701
mag
−0.15
−0.25
information gain
P(correct): 0.994
Entropy: 0.04
0
5
10
15
20
25
t

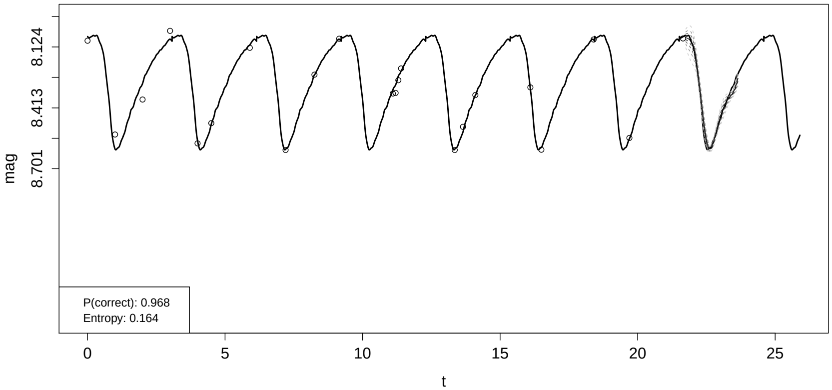

mag
8.124
8.413
8.701
0
5
10
15
20
25
t
P(correct): 0.98
Entropy: 0.112

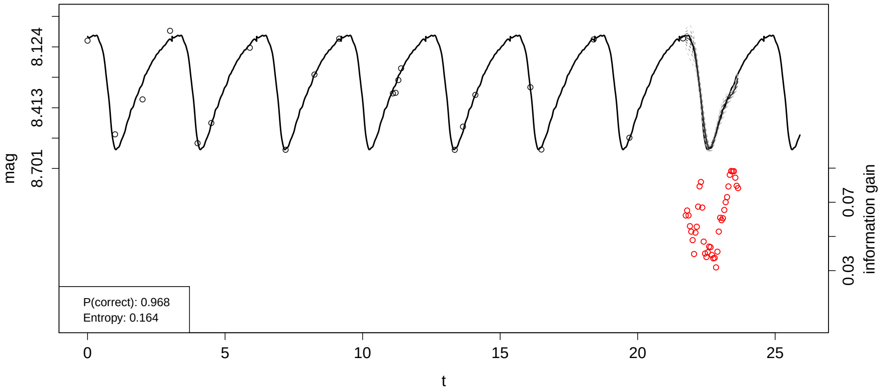

mag
8.124
8.413
8.701
0
5
10
15
20
25
t
P(correct): 0.98
Entropy: 0.112

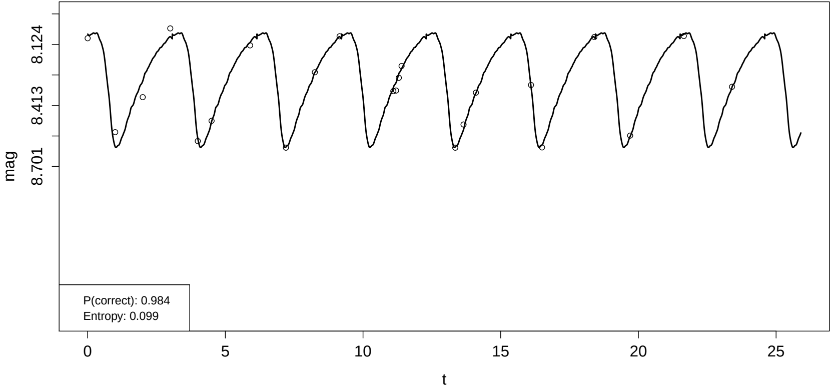

8.124
8.413
8.701
mag
0.00
−0.05
information gain
P(correct): 0.98
Entropy: 0.112
0
5
10
15
20
25
t



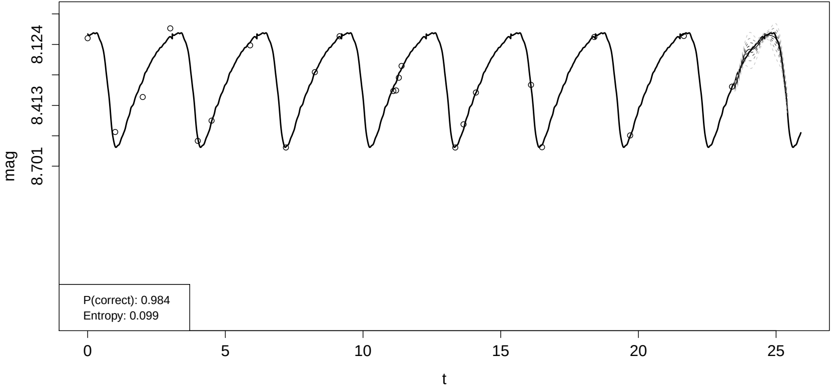
mag
8.124
8.413
8.701
0
5
10
15
20
25
t
P(correct): 0.968
Entropy: 0.164

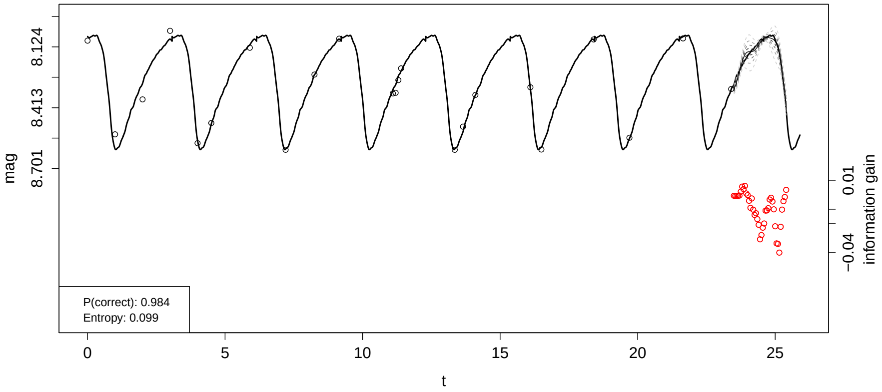

mag
8.124
8.413
8.701
P(correct): 0.968
Entropy: 0.164
0
5
10
15
20
25
t

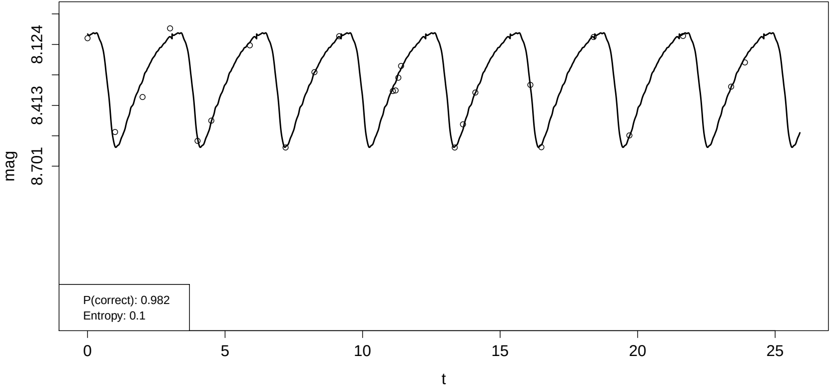

mag
8.124
8.413
8.701
information gain
0.07
0.03
P(correct): 0.968
Entropy: 0.164
0
5
10
15
20
25
t

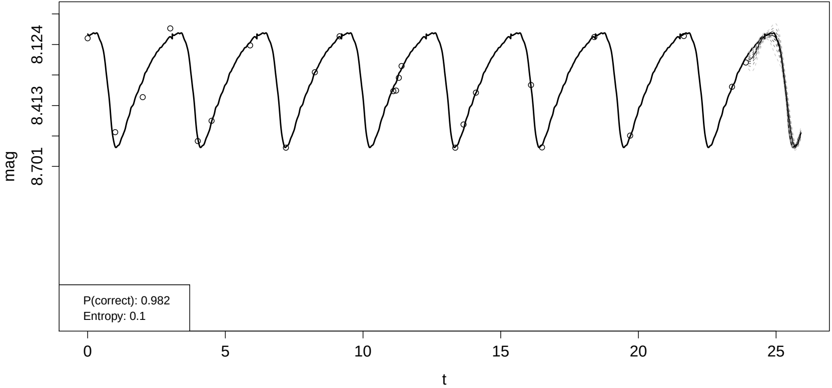


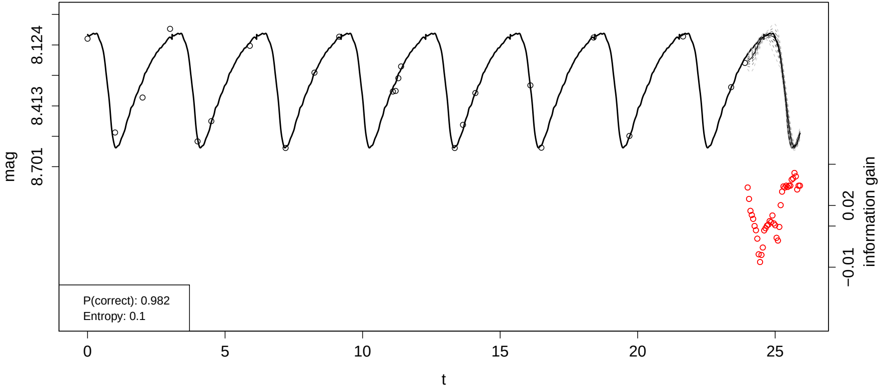
P(correct): 0.984
Entropy: 0.099
mag
8.124
8.413
8.701
0
5
10
15
20
25
t


mag
8.124
8.413
8.701
0
5
10
15
20
25
t
P(correct): 0.984
Entropy: 0.099


mag
8.124
8.413
8.701
0.01
−0.04
information gain
P(correct): 0.984
Entropy: 0.099
0
5
10
15
20
25
t


mag
8.124
8.413
8.701
0
5
10
15
20
25
t
P(correct): 0.982
Entropy: 0.1


8.124
8.413
8.701
mag
P(correct): 0.982
Entropy: 0.1
0
5
10
15
20
25
t


P(correct): 0.982
Entropy: 0.1
mag
8.124
8.413
8.701
0
5
10
15
20
25
t
information gain
0.02
−0.01