A Latent Source Model for Nonparametric Time Series Classification

George H. Chen § Stanislav Nikolov † Devavrat Shah §

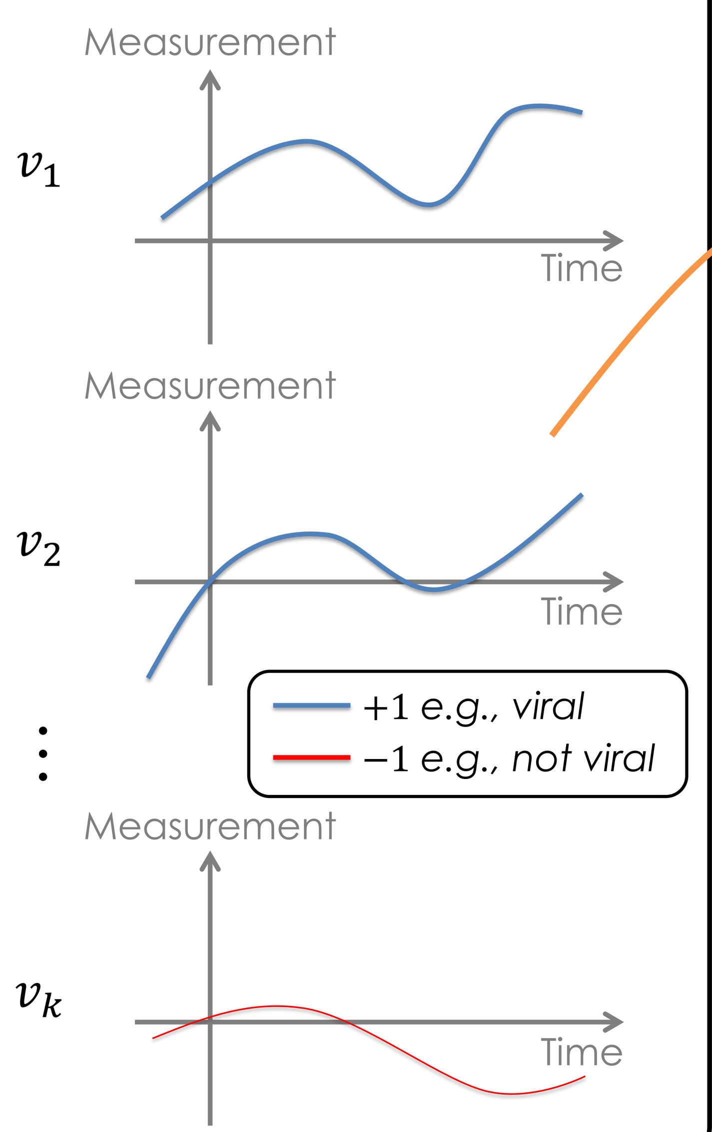
Motivation

- Nearest-neighbor-like methods for time series classification widely used in practice, often with outstanding performance
- Little theoretical development for characterizing performance in terms of:
  → How much training data to use?
  → How much of time series do we observe?

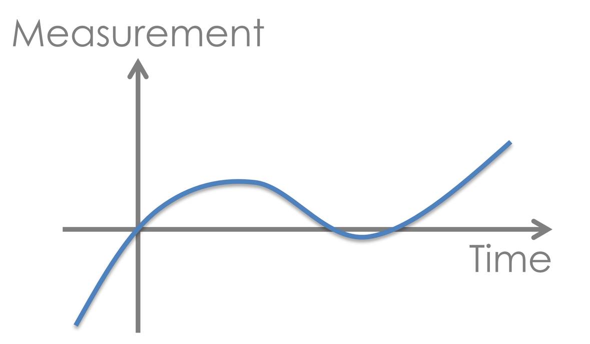
Goal: develop theory to explain performance of nearest-neighbor-like methods for time series classification, and relate theory to practice (*forecasting which news topics will go viral on Twitter*)

Hypothesis: in many real time series datasets, there are only a *few* possible patterns (*latent sources*) relative to how many time series we can collect (*a news topic goes viral on Twitter only in a few ways yet we can collect time series for a huge number of news topics*)

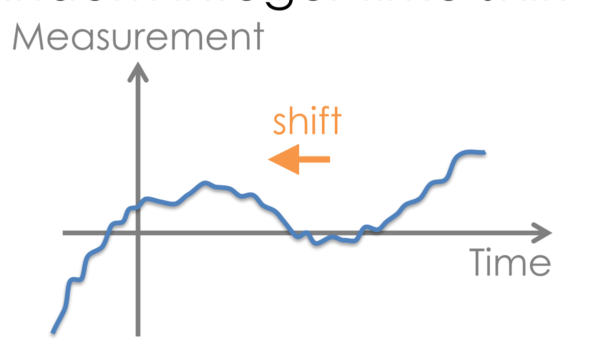
A Latent Source Model



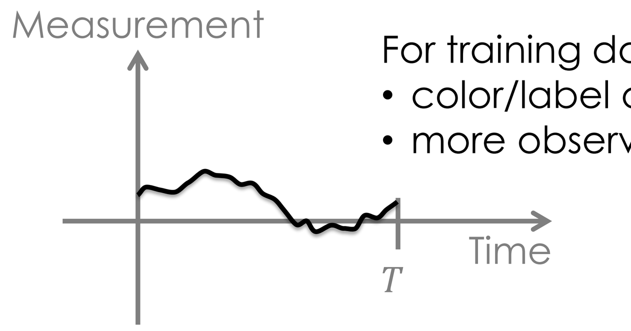
1. Choose random latent source



2. Add i.i.d. zero-mean sub-Gaussian noise & random integer time shift $\{0,1,\dots,\Delta_{\max}\}$



3. Observe time steps $1,2,\dots,T$



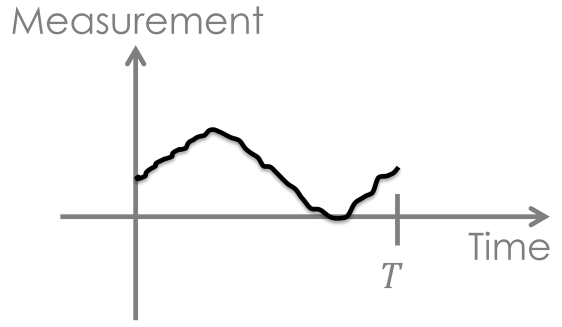
For training data, also get:
- color/label of time series
- more observed time steps

For test data: want to determine color/label

Binary Classification

Oracle MAP estimator (if noise is Gaussian and we knew the latent sources)

Time series s to be classified



Compute ℓ_2 distance to time shifted versions (denoted $v_i * \Delta$) of each latent source v_i masked to time steps $1,2,\dots,T$

$$\|s - v_i * \Delta\|_T^2 \triangleq \sum_{t=1}^{T} \left(s(t) - v_i(t+\Delta)\right)^2$$

Likelihood ratio test:

$$\frac{\sum_{v_i \in V_+} \sum_\Delta \exp\left(-\frac{1}{2\sigma^2}\|s - v_i * \Delta\|_T^2\right)}{\sum_{v_i \in V_-} \sum_\Delta \exp\left(-\frac{1}{2\sigma^2}\|s - v_i * \Delta\|_T^2\right)} \overset{?}{\geq} 1$$

If true, declare label of s to be +1

We don't actually know the latent sources!
→ use training data $r_1, r_2, \dots, r_n$ generated from latent source model as proxy (assume each is observed for all time and come with ground truth labels)

Weighted majority voting
Approximation of oracle MAP

$$\frac{\sum_{r_i \in R_+} \exp\left(-\gamma \min_\Delta \|s - r_i * \Delta\|_T^2\right)}{\sum_{r_i \in R_-} \exp\left(-\gamma \min_\Delta \|s - r_i * \Delta\|_T^2\right)} \overset{?}{\geq} 1$$

If true, declare label of s to be +1

Nearest neighbor classifier
Approximation of weighted majority voting

$$\hat{r} = \underset{r \in R_+ \cup R_-}{\operatorname{argmin}} \min_\Delta \|s - r_i * \Delta\|_T$$

Declare label of s to be the same as that of $\hat{r}$

Theorem: Under the latent source model, with $n = \Theta\left(k \log \frac{k}{\delta}\right)$ training time series, if

(gap) $$\min_{\substack{r_+ \in R_+, r_- \in R_-, \\ \Delta_+, \Delta_-}} \|r_+ * \Delta_+ - r_- * \Delta_-\|_T^2 = \Omega(\sigma^2 T),$$

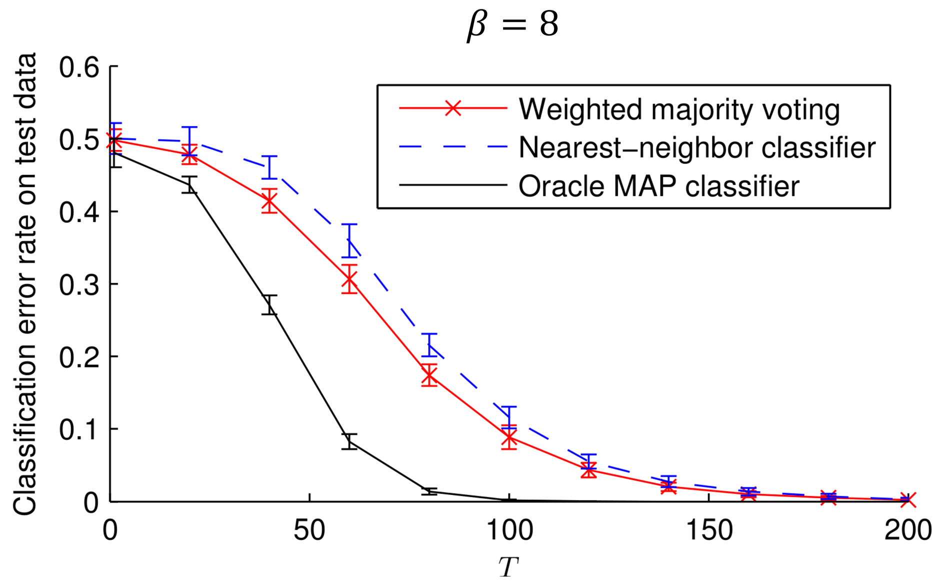
then weighted majority voting (with $\gamma = \frac{1}{8\sigma^2}$) and nearest-neighbor classification each classify time series s correctly with probability at least $1 - \delta$ once we've seen the first $T = \Omega\left(\log(2\Delta_{\max} + 1) + \log \frac{k}{\delta}\right)$ time steps of s.

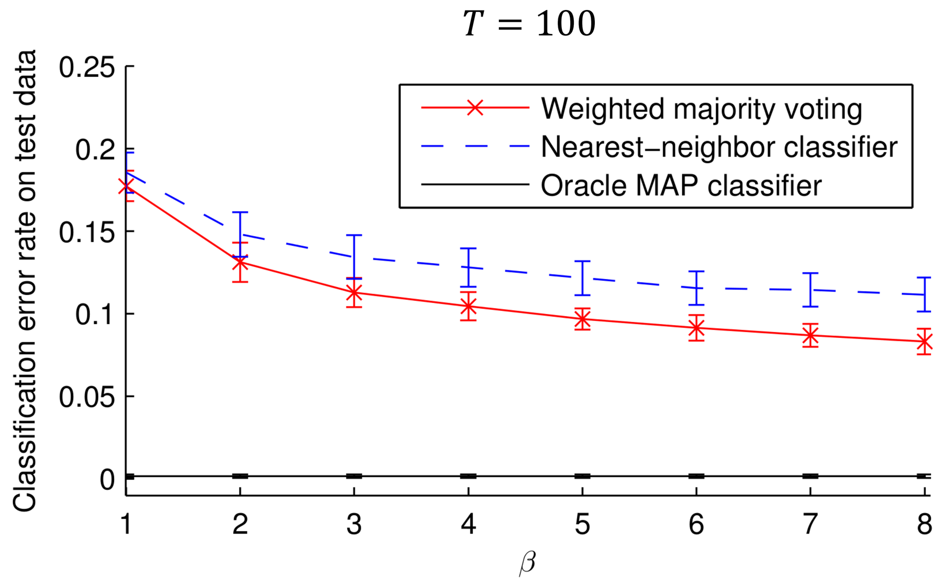
Why not just learn the latent sources?
→ For Gaussian noise and no time shifts, existing results on learning Gaussian mixture models require more training data than what our results require or require more stringent assumptions on separation of mixture components

Experimental Results

Synthetic data

- $k = 200$ latent sources, ½ with label +1, ½ with label −1; generate each from Gaussian process: $\mathcal{N}(0,100)$ smoothed w/ 1D Gauss filter (scale 30)
- Noise is $\mathcal{N}(0,1)$, max shift is $\Delta_{\max} = 100$
- Sample $n = \beta k \log k$ training time series



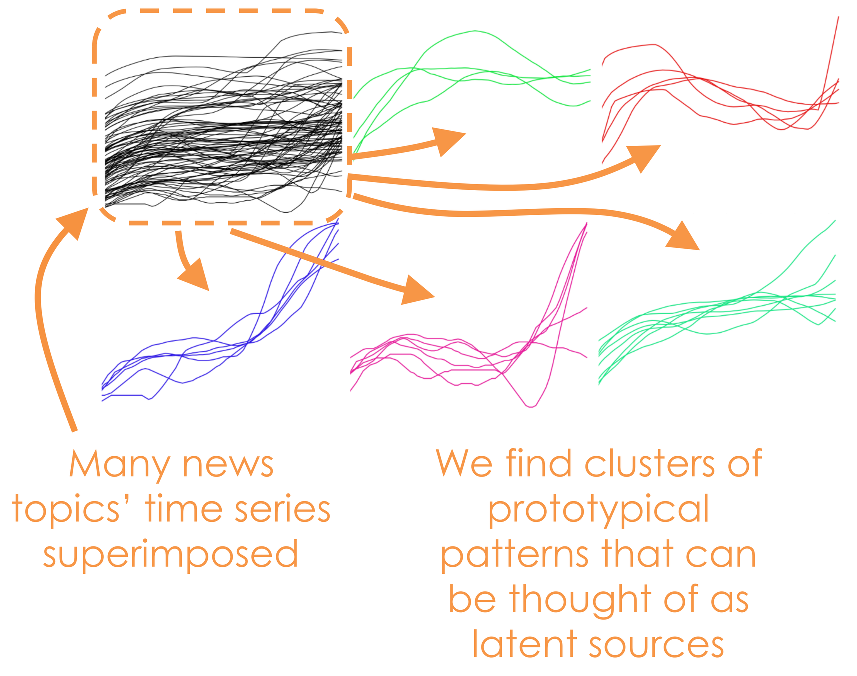


- Weighted majority voting ($\gamma = 1/8$) and nearest-neighbor classification have similar performance for large T in agreement with theory
- For small T, weighted majority voting outperforms nearest-neighbor classification → weighted majority voting better suited than NN classification for *online* time series classification

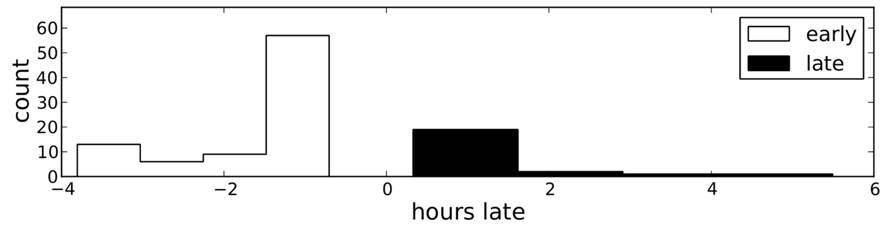
Forecasting which news topics will go viral on Twitter

- 500 new topics that go viral; 500 that do not → split 50%-50% train-test
- Forecast trends on test dataset
- Online time series classification

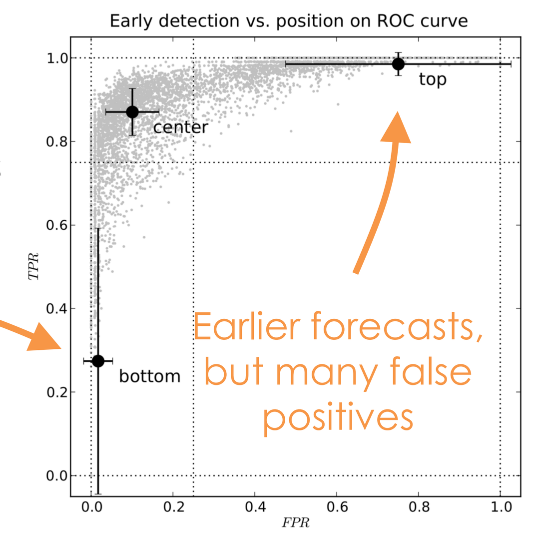
How news topics go viral on Twitter



Weighted majority voting finds which news topics will go viral before Twitter does 79% of the time, with an early advantage of 1.43 hours; FPR: 4%, TPR: 95%



ROC curve guides choice of parameter settings



This work was supported in part by the Army Research Office under MURI Award 58153-MA-MUR. GHC was supported by an NDSEG fellowship.