Feature Flocks Derivation

Davis W. Blalock & John V. Guttag
MIT CSAIL

In this document, we describe the theoretic basis for Feature Flocks. To do so, we first consider a concrete and simpler problem in order to build the reader's intuition, and then detail how this can be applied to time series.

I. The Robbers and Randoms Problem

A. Intuition

Suppose that a gang of robbers has stolen something from your institution. Police have identified a group of suspects, but this group contains both robbers and random civilians (drawn iid from the general population in your city). As the local data scientist, it is up to you to determine who the robbers are. You cannot assume a certain number of robbers or anything about their characteristics (e.g., they do not necessarily have criminal records or other obvious features to look for). You can assume, however, that there are many more robbers in the group than would be expected by chance if you only sampled iid from the general population.

A priori, this is not necessarily solvable. If the robbers have nothing in particular that distinguishes them, then there can be little hope of identifying them. However, it is probable that fellow robbers will have some sort of distinguishing features. For example, if several of the suspects all happen to have played on the same football team in high school, it is unlikely that this would happen in iid civilians, and so you might suspect that these are the robbers. If you can determine what these features are, then, you can recognize the robbers. So how could you identify these features? The above example seems sensible enough, but what is it about a given feature that makes it useful?

One rule is that the feature must be *distinguishing*—i.e., not common in the general populace. "Has brown hair" is probably a poor feature.

However, it is not enough for a feature to be unusual. Indeed, any person in the group is likely to have a number of features that are not just unusual, but unique. We cannot simply lock up, say, the one redhead.

However, if multiple people have the *same* unusual feature, we might suspect that this is not a coincidence. For example, if three of our suspects happen to be from the same neighborhood, or work for a particular company, this is unlikely under the assumption that they're innocents drawn from the general population. This becomes even more unlikely the more suspects there are who have this trait in common—two people may work for a certain company by chance, but five almost certainly do not.

Further, if this same set of individuals has other features in common (e.g., graduating from the same high school, being the same age), this bolsters our conviction.

In short, we seek features with three properties:

1. **rare** in general
2. **common** in a certain subset of the people (whom we suspect are the robbers)
3. shared across **many** (suspected) robbers

B. Objective Function

Formally, you are given a set X of feature descriptions of individuals, x_i, where x_i^j is the jth feature of the ith individual. Your task is to return the most probable set of robbers, I^*, and distinguishing features, F^*. There are N individuals, at least two robbers ($|I| \geq 2$) and $D = |x_i|$ features. Your objective function is defined as follows:

$$I^*, F^* = \arg\max_{I,F} \prod_{i \in I} \prod_{j \in F} \frac{p(x_i^j | z_i = 1)}{p(x_i^j | z_i = 0)} \tag{1}$$

where z is an indicator variable describing whether $i \in I$, and different distributions are associated with each value of z. In words, we want to find a set of individuals and features such that the probability of getting all of those features across all of those individuals is much higher if the individuals are pulled from a shared "robber" distribution than the "random" general populace distribution. We will describe the family of distributions we use in solving our particular problem in following sections, but all that is required for the above objective is that the "robber" and "random" distributions not be identical.

C. Objective Function Derivation

The above objective function yields the maximum *a posteriori* (MAP) estimate of the set of robbers and their distinguishing features given a few simple assumptions.

By definition, the most probable set of objects and features is given by:

$$\arg\max_{I,F} p(I, F | X) \tag{2}$$

Using Bayes Theorem and dropping the denominator (since it does not affect the argmax), we see that this is equivalent to:

$$\arg\max_{I,F} p(X | I, F) p(I, F) \tag{3}$$

Now suppose that our priors regarding the set of people and set of features are independent. We then have:

$$\arg\max_{I,F} p(X | I, F) p(I) p(F) \tag{4}$$

The first of these probabilities, $p(X|I,F)$ can be simplified further given a few reasonable assumptions. First, since we consider each object to have been drawn iid, we have:

$$p(X|I,F) = \prod_{i=1}^{N} p(x_i|I,F) \tag{5}$$

Now, we introduce the latent indicator variable z, where $z_i = 1 \iff i \in I$, to encapsulate the conditioning of x_i on I. This allows us to rewrite the previous equation as:

$$p(X|I,F) = \prod_{i\in I} p(x_i|F,z=1) \prod_{i\notin I} p(x_i|F,z=0) \tag{6}$$

$$= \prod_{i\in I} \frac{p(x_i|F,z=1)}{p(x_i|F,z=0)} \prod_{i=1}^{N} p(x_i|F,z=0) \tag{7}$$

where we have multiplied in $\prod_{i\in I} p(x_i|F,z=0)$ to get the second equation. Now let us define:

$$p(x_i|F,z=0) = p(x_i|z=0) \tag{8}$$

This reflects the idea that the probability of getting x_i from the "random" distribution is not dependent on which features we select as relevant for the "robber" distribution. E.g. if we determine that the robbers are likely to be from New York, this has no bearing on the probability that a random person is from New York. This definition renders the second product in (6) independent of both I and F, allowing us to drop it without affecting our original argmax:

$$p(X|I,F) \propto \prod_{i\in I} \frac{p(x_i|F,z=1)}{p(x_i|z=0)} \tag{9}$$

We can reasonably make another assumption to simplify this even further. If only the features $j \in F$ are determined by an object's status as a robber, and others are determined by whatever processes produce the features of the general population, then one could reasonably factor the overall probability of an object into the probability of the robber and non-robber features. That is, we can define:

$$p(x|F,z) \equiv p(x^F|z)p(x^{-F}|z) \tag{10}$$

$$p(x^{-F}, z=1) \equiv p(x^{-F}|z=0) \tag{11}$$

where x^F and x^{-F} are the features of x in and outside of the set F, respectively. I.e., $x^F = \langle x^j \rangle, j \in F$ and $x^{-F} = \langle x^j \rangle, j \notin F$. Intuitively, this means that if brown hair has nothing to do with being a robber, then the probability of having brown hair is the same for robbers and non-robbers. Further, the relevant and irrelevant features are independent of one another conditioned on whether an individual is a robber. Using these properties, we can simplify 9 into:

$$p(X|I,F) \propto \prod_{i\in I} \frac{p(x_i^F|z=1)}{p(x_i^F|z=0)} \frac{p(x_i^{-F}|z=1)}{p(x_i^{-F}|z=0)} \tag{12}$$

$$= \prod_{i\in I} \frac{p(x_i^F|z=1)}{p(x_i^F|z=0)} \tag{13}$$

where we have vacuously conditioned on F in the denominator of 9 in order to obtain 12.

Now consider what happens if we assume that $p(x_i^F|z)$ factorizes—i.e., that the individual features are independent when conditioned on z. This gives us:

$$p(X|I,F) \propto \prod_{i\in I} \prod_{j\in F} \frac{p(x_i^j|z=1)}{p(x_i^j|z=0)} \tag{14}$$

This yields the final objective function:

$$I^*, F^* = \arg\max_{I,F} p(I)p(F) \prod_{i\in I} \prod_{j\in F} \frac{p(x_i^j|z=1)}{p(x_i^j|z=0)} \tag{15}$$

If we assume uniform priors $p(I)$ and $p(F)$, so that these terms do not affect the argmax, we recover the original objective function in 1. Thus, this objective is a MAP estimate (or, ignoring the priors entirely, a maximum likelihood estimate).

D. Binary Case

The above analysis works for any distributions satisfying the stated independence assumptions. To identify the robbers, however, we must select distributions to fit to the data. Throughout the rest of this work, we will use $Bernoulli$ distributions and binary features.

Specifically, we define $p(x_i^j|z=0) \sim Bernoulli(\theta_{0j})$ and $p(x_i^j|z=1) \sim Bernoulli(\theta_{1j})$, where $\boldsymbol{\theta_0}$ is any set of parameters computed independent of I and F, and $\boldsymbol{\theta_1}$ is the optimal set of parameters (in this case, simply the maximum likelihood estimate) for $\{x_i^j|i \in I\}$. We also drop $p(F)$ and $p(I)$ for now—the latter will be relevant when we generalize to time series. Taken together, these changes give the objective function:

$$I^*, F^* = \arg\max_{I,F} \prod_{i\in I} \prod_{j\in F} \frac{p(x_i^j|\theta_{1j})}{p(x_i^j|\theta_{0j})} \tag{16}$$

Expanding the two cases for each feature, we have:

$$\arg\max_{I,F} \prod_{i\in I} \prod_{j\in F} \frac{\theta_{1j}^{I\{x_i^j=1\}}(1-\theta_{1j})^{I\{x_i^j=0\}}}{\theta_{0j}^{I\{x_i^j=1\}}(1-\theta_{0j})^{I\{x_i^j=0\}}} \tag{17}$$

Letting c_j denote the number of times feature j is 1, $\sum_{i\in I} x_i^j$, and k denote $|I|$, this can be rewritten as:

$$\arg\max_{I,F} \prod_{j\in F} \frac{\theta_{1j}^{c_j}(1-\theta_{1j})^{(k-c_j)}}{\theta_{0j}^{c_j}(1-\theta_0 j)^{(k-c_j)}} \tag{18}$$

Now, let us take the log of this objective function (which yields the same answer, since the log is a monotonic function of its argument):

$$\arg\max_{I,F} \sum_{j\in F} [c_j(log(\theta_{1j}) - log(\theta_{0j})) \tag{19}$$

$$+(k-c_j)(log(1-\theta_{1j}) - log(1-\theta_{0j}))] \tag{20}$$

Another simplification is possible if the data is sparse. In this case, it is not meaningful for a feature to be absent. E.g., if a feature is only present in 10% of the population, the fact that it is not present in three robbers is not a robust indication that being a robber affects it. For example, if none of the three robbers have been to Greenland, we should not take failure to visit this country as a sign of being a robber. This is crucial, since one could enumerate countless features of this nature and incorporating them would almost certainly drown out the features that were meaningful. Dropping the term for feature absence in the above equation, we have:

$$\arg\max_{I,F} \sum_{j \in F} c_j(log(\theta_{1j}) - log(\theta_{0j})) \tag{21}$$

This equation says that we would like to find robbers x_i and features j such that x_i^j happens both many times (so that c_j is large) and much more often than would occur by chance (so that $log(\theta_{1j}) - log(\theta_{0j})$ is large). This is the objective function given in the problem definition section.

E. Relationship to compression

It is interesting to note that the binary objective connects directly to the problem of compression. Suppose that $\boldsymbol{\theta_0}$ is set to the emprical probabilities across the whole dataset, so that $\theta_{0j} = p(x^j)$. Suppose further than we can only select features that that are always present, so that $c_j = |I|$ and $\boldsymbol{\theta_1} = \mathbf{1}$. Then the objective becomes:

$$\arg\max_{I,F} \sum_{j \in F} -|I| * \log(\theta_{0j}) \tag{22}$$

$$= \arg\max_{I,F} \sum_{j \in F} -|I| * \log(p(x^j)) \tag{23}$$

This is the number of bits to encode k occurrences of all features j, with codeword length determined by the empirical probability $p(x^j)$. In other words, if each feature j would otherwise be encoded at a cost of $\log(p(x^j))$ when present, this is the number of bits saved by substituting for 1s in $x_i^F, i \in I$ a symbol composed of ones at the indices F. Thus, this objective is maximized when we find the largest set of ones that occur together the most times throughout the data.

II. Generalizing to multiple models

The analysis so far has considered only distinguishing between two models: a "robber" model and a "random" model. This can be generalized to many models. Specifically, while we are still interested only in identifying the robbers, non-robbers may be drawn from many distributions, rather than a naive aggregate one. This is a more realistic setup, but a more challenging one mathematically. It requires us to ensure that suspected robbers not only fail to appear random, but fail to resemble any competing distribution as well.

Formally, let $\mathcal{M} = \{\mathbf{m}_l\}$ be some set of possible models, and $\boldsymbol{\theta_1}$ continue to be the robber model. For ease of reference, further define $X_I \equiv \{x_i, i \in I\}$ and $X_{-I} \equiv \{x_i, i \notin I\}$. We seek to find:

$$\arg\max_{I,F} p(I)p(F) \frac{p(X|I,F,\boldsymbol{\theta_1})}{\sum_{\mathbf{m} \in \mathcal{M}} p(X|I,F,\mathbf{m})p(\mathbf{m})} \tag{24}$$

$$= \arg\max_{I,F} p(I)p(F) \prod_{i \in I} \frac{p(x_i|I,F,\boldsymbol{\theta_1})}{\sum_{\mathbf{m} \in \mathcal{M}} p(x_i|I,F,\mathbf{m})p(\mathbf{m})} \tag{25}$$

Since there can be many models, this quantity is challenging to optimize. Therefore, we will approximate it as follows, using the intuition that one model likely assigns X_I much greater probability than the others:

$$\arg\max_{I,F} \min_{\mathbf{m} \in \mathcal{M}} p(I)p(F) \prod_{i \in I} \frac{p(x_i|I,F,\boldsymbol{\theta_1})}{p(x_i|I,F,\mathbf{m})} \tag{26}$$

That is, we approximate the sum with the maximum element and set its prior $p(\mathbf{m}) = 1$.

We must now define the set $\mathcal{M}$. One could use any number of possible sets of models, but we will use a simple instance-based approach. In our robber-finding example there are reasonable alternatives, since the general populace is large and one could learn a set of models for it. However, when generalizing to time series, we will assume extremely limited data, and so it will be impossible to learn a set of parametric distributions (or even how many such distributions there should be) with accuracy.

Concretely, we take each object X_{-I} as the "centroid" of a model $\mathbf{m}$. This means that the probability of a suspected robber x_i coming from a distribution other than the robber distribution is based on its similarity to the most similar non-robber object. For ease of reference, we term this object the *nearest enemy*, and the suspected robber x_i the *query*.

But what defines this similarity, and do we get a probability from it? One approach would be to have a kernel function that, say, returned a probability between 0 and 1 depending on the Hamming distance between the query and nearest enemy. This could work, but we would have to learn the mapping of distances to probabilities. We would also consider deviations in all features equal (by definition of the Hamming distance), while in reality, some features may be more variable than others.

What we do instead is recycle the learned distribution for the robbers, described by θ_1. Specifically, letting x_k denote the nearest enemy, we define:

$$p(x_i|I,F,\mathbf{m}_k) \equiv p(x_k|I,F,\boldsymbol{\theta_1}) \tag{27}$$

In words, we approximate the probability of the query coming from the distribution of the neareset enemy with the probability of the nearest enemy coming from the distribution of the query (i.e., the robber distribution). This yields the objective:

$$\arg\max_{I,F} \min_{k \notin I} p(I)p(F) \prod_{i \in I} \frac{p(x_i|I,F,\boldsymbol{\theta_1})}{p(x_k|I,F,\boldsymbol{\theta_1})} \tag{28}$$

$$\tag{29}$$

Taking the log, this becomes:

$$\begin{aligned}\arg\max_{I,F}\min_{k\notin I}&\log(p(I))+\log(p(F))\\ &-|I|\log(p(x_k|I,F,\theta_{\mathbf{1}}))\\ &+\sum_{i\in I}\log(p(x_i|I,F,\theta_{\mathbf{1}}))\end{aligned}$$

In the binary case (using the previously described approximations and independencies), we have:

$$\begin{aligned}\arg\max_{I,F}\min_{k\notin I}&\log(p(I))+\log(p(F))\\ &+\sum_{j\in F}(c_j-|I|x_k^j)\log(\theta_{1j})\end{aligned}$$

Unfortunately, this objective is degenerate—since $\log(\theta_{1j})$ is always negative, it is maximized when $I=F=\{\}$. Thus, we instead maximize:

$$\arg\max_{I,F}\min_{k\notin I}\log(p(I))+\log(p(F)) \qquad (30)$$

$$+\sum_{j\in F}(c_j-|I|x_k^j)\log(\theta_{1j}-\theta_{0j}) \qquad (31)$$

In other words, we continue looking for features that are better explained by the robber distribution than the noise distribution, but exclude those that are better explained by the nearest enemy distribution.

To handle the case when all objects are suspected of being robbers (and thus there is no object eligible to be the enemy), we add to the dataset a "dummy" enemy x_k such that $x_k^j=E[\theta_{\mathbf{0}}]$. This object represents the "expected" enemy given features drawn from the "noise" distribution. Empirically, this object may also be chosen as the nearest enemy if the other possible enemies are extremely sparse or otherwise share few features with the suspected robbers.

III. Generalizing to Time Series

The above objective and analysis can be applied to time series via a straightforward reduction. Namely, instead of having each object x_i be a binary feature representation of an individual, it is a binary feature representation of a region (a,b) of the time series. This makes no difference mathematically—it simply changes what each object represents.

However, one mathematical alteration is necessary for the results to be meaningful. Namely, we use the prior $p(I)$ to prevent overlapping, excessively long, or excessively short regions. That is:

- Given a minimum spacing M_{space} between the starts of regions, $p(I)=0$ if $|a_1-a_2|<M_{space}$ for any regions $(a_1,b_1),(a_2,b_2)\in I$.
- Given minimum and maximum region lengths M_{min} and M_{max}, $p(I)=0$ if $|b_1-a_1|<M_{min}\vee|b_1-a_1|>M_{max}$ for any $(a_1,b_1)\in I$.
- $p(I)$ is otherwise uniform.

This yields the final objective (without enemies):

$$\arg\max_{I,F}\log(p(I))+\sum_{j\in F}c_j(log(\theta_{1j})-log(\theta_{0j})) \qquad (32)$$

With enemies, this becomes:

$$\arg\max_{I,F}\min_{k\notin I}\log(p(I))+\sum_{j\in F}(c_j-|I|x_k^j)\log(\theta_{1j}-\theta 0j) \qquad (33)$$

A. Flock Filters

In the time series case, maximizing the first of the above objectives (32) can be interpreted as learning a digital filter that selectively responds to pattern instances.

Let $\mathbf{w}=V(log\theta_{\mathbf{1}}-log\theta_{\mathbf{0}})$, where V is a binary diagonal matrix that zeros out all the features that aren't included in F. Then $\mathbf{w}^\top\mathbf{x}$ is the difference in log likelihoods of getting $\mathbf{x}$ under the "pattern" distribution and "non-pattern" distribution, which is also the log odds of it being a pattern instance. If the feature representation x_i is taken not as a vector but a 2D window, $\mathbf{w}$ is a 2D digital filter that can be slid along the time dimension to assess each window of data. This is the mathematical meaning of the "Learned Pattern" in Figure 8 of the paper.