Semi-Parametric Manifold Clustering

Estimating Polynomial Curves

Problem Setup

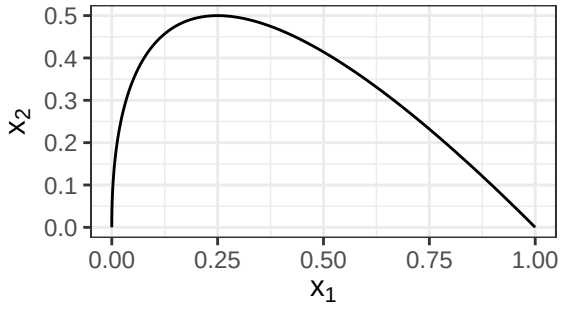
Let:

- $T_1, ..., T_n \overset{\text{iid}}{\sim} F$ with support $[0, 1]$.
- $g(\cdot, \theta) : [0, 1] \mapsto \mathcal{X} \subset \mathbb{R}^d$.
- $X_1, ..., X_n = g(T_1), ..., g(T_n)$

Assuming some parametric form of g with parameters θ, we want to find $\hat{\theta}$, some "reasonable" estimate for θ. We observe X_i but not T_i.

For now, we limit $d = 2$ and g to quadratic functions.

Example 1. *Let $g(t) = (t^2, 2t(1-t)) = (t^2, 2t - 2t^2)$. (This is the first two dimensions of the Hardy-Weinberg curve). Then $\theta = (0, 0, 1, 0, 2, -2)$.*



If we observe the T_i's, then we can use a standard polynomial regression method to obtain $\hat{\theta}$. Since we do not observe them, the proposed iterative method is as follows:

1. Initialize $\hat{\theta}^{(0)}$ (e.g., by drawing from a probability distribution).
2. Estimate each $\hat{t}_i^{(s)}$ by minimizing $L(t_i, \hat{\theta}^{(s)}|x_i) = L_i = \|x_i - g(t_i|\hat{\theta}^{(s)})\|^2$.
3. Compute each $\hat{x}_i^{(s)} = g(\hat{t}_i^{(s)}|\hat{\theta}^{(s)})$
4. Estimate $\hat{\theta}^{(s+1)}$ by minimizing $L(\{\hat{t}_i^{(s)}\}, \theta|X) = \sum_i \|x_i - g(\hat{t}_i^{(s)}|\theta)\|^2$.
5. Repeat steps 2-4 until convergence.

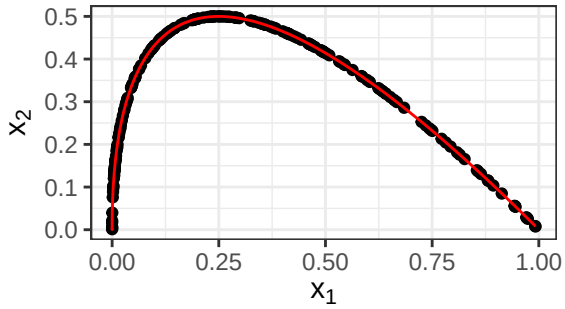
If we restrict g to be polynomials, then steps (2) and (4) have closed-form solutions.

Example 2. *Write $g(t|\theta) = (g_1(t|\theta_1), ..., g_d(t|\theta_d))$ where $g_r(t|\theta_r)$ is the component of g in the r^{th} dimension and θ_r is the vector of parameters for the r^{th} dimension. If g_r are polynomials of degree p, then each θ_r contains up to $p + 1$ entries.*

Given the observed points $x_1, ..., x_n \in \mathbb{R}^d$ and their corresponding index points $t_1, ..., t_n \in \mathbb{R}$, we can find each $\hat{\theta}_r$ individually by $\hat{\theta}_r = A^{-1}b$ where $b \in \mathbb{R}^{p+1}$ and $b_k = \sum_i x_i t_i^k$ and $A \in \mathbb{R}^{(p+1)\times(p+1)}$ and $A_{kl} = \sum_i t^{(k-1)(l-1)}$.

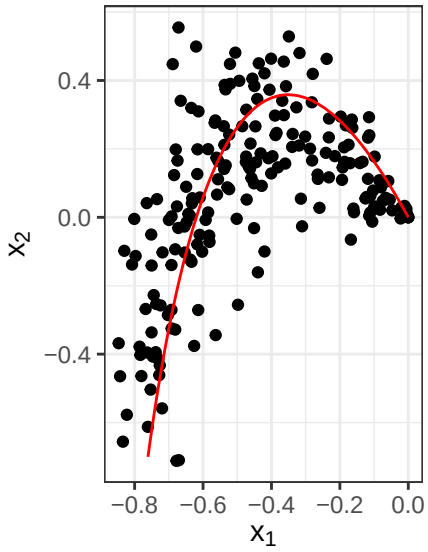
On the other hand, if we have parameters θ but not the index points t_i, we can minimize each t_i individually by finding the roots of a $p+1$ polynomial with coefficients that depend on $x_1, ..., x_n$ and θ.

In the following plot, we drew $n = 200$ points from the 2D H-W curve with $T_1, ..., T_n \overset{\text{iid}}{\sim} Uniform(0, 1)$. The red line is the curve that was fit using the above method.



One problem with this method is the parameterization of the curve is not unique.

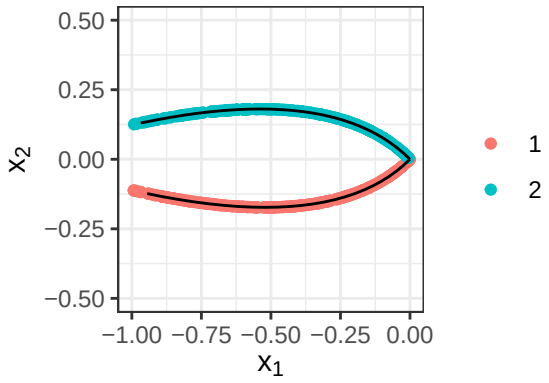
Example 3. *In the next example, we draw $A \sim \text{RDPG}(X)$ using the same H-W curve and sample size as above and estimate the true latent positions (up to rotation). In this example, we force the intercept terms to be zero.*



Next, suppose we have K curves parameterized by $g^{(k)}$, with points drawn along these curves. Then one possible clustering technique is as follows:

1. Assign an initial clustering (e.g., via spectral clustering).
2. Estimate the curve for each cluster.
3. Reassign the clusters by proximity to each cluster.
4. Repeat 2 and 3 until convergence.

Example 4. *We again limit these to be quadratic functions in $\mathbb{R}^2$. Here, $K = 2$ and $n_1 = n_2 = 256$.*



Example 5. *Finally, we draw* $A \sim \mathrm{RDPG}(X)$ *from the above example and apply the clustering and curve fitting to the ASE of* A*.*

```
   zhat.manifold
z      1   2
  1 187  69
  2  87 169
```