Notes on Nonparametric Partial Identification of Mixtures

Maxim Rabinovich

March 23, 2017

1 Problem setting and overview

We are interested in the simple d-dimensional mixture model

$$\begin{aligned} Z &\sim \text{Bern}(\pi), \\ X \mid Z = z &\sim \mathbb{P}_z, \end{aligned}$$

where $\mathbb{P}_z$ is some distribution on $\mathbb{R}^d$. We assume π is known to the practitioner, but that the $\mathbb{P}_z$ are not and that estimating them is not desired. Rather, we suppose the practitioner seeks to identify the parameters:

$$\mu_z = \mathbb{E}_{\mathbb{P}_z}[X], \quad z \in \{0,\ 1\}.$$

Without either (or both) parametric assumptions on $\mathbb{P}_z$ or independence assumptions between the dimensions of X, full identification is not possible in general. We therefore settle for partial identification, in the sense that we seek a valid α-confidence region $\mathcal{M} \subset \mathbb{R}^d \times \mathbb{R}^d$ for $\mu_{0:1} \in \mathbb{R}^d \times \mathbb{R}^d$.

The general approach we take is as follows. Let $\mathcal{P}$ be some collection of allowed pairs $(\mathbb{P}_0,\ \mathbb{P}_1)$. This set could be parametric—e.g. all Gaussians—but the case that interests us is when $\mathcal{P}$ is entirely nonparametric and defined only in terms of structural constraints on the moments of the $\mathbb{P}_z$. The class could be defined by the structure of the covariance matrix $\mathbb{E}\left[XX^T\right]$, or by an assumption on the size of $||\mu_1 - \mu_0||_\infty$.

Given a class of candidates $\mathcal{P}$, we then consider a class $\mathcal{F}$ of functions that we call *moment probes*. This class simply describes all moments of the distribution that we shall constrain to be close to the observed values. The most naive way to generate a partial identification $\mathcal{M}$ given the information we have so far is to solve the following abstract feasibility program:

$$\begin{aligned} \text{exists } (\mathbb{P}_0,\ \mathbb{P}_1) \quad \text{s.t.} \quad & \\ & (\mathbb{P}_0,\ \mathbb{P}_1) \in \mathcal{P}, \\ & \forall f \in \mathcal{F}, \quad \pi f(\mathbb{P}_0) + (1-\pi) f(\mathbb{P}_1) = f\left(\hat{\mathbb{P}}\right), \end{aligned} \tag{1}$$

where $\hat{\mathbb{P}}$ is the empirical distribution of the observed X_i's and $f(\mathbb{P}) = \mathbb{E}_{\mathbb{P}}[f(X)]$. The solution to the feasibility program (1) is a set $\hat{\mathcal{P}} \subset \mathcal{P}$ and returning $\mathcal{M} = \left\{\mu_{0:1} \colon \mu_z = \mathbb{E}_{\mathbb{P}_z}[X], (\mathbb{P}_0,\ \mathbb{P}_1) \in \hat{\mathcal{P}}\right\}$ would be a partial identification—though not necessarily a valid one, as it does not account for estimation error.

In order to ensure validity, at the cost of some power, we may suppose given $\epsilon_f,\ f \in \mathcal{F}$ such

that $\prod_{f\in\mathcal{F}}\left[f\left(\hat{\mathbb{P}}\right)\pm\epsilon_f\right]$ is a valid α-confidence region for $\left(f\left(\mathbb{P}\right)\right)_{f\in\mathcal{F}}$. We can then modify (1) into

$$\begin{aligned}\text{exists } (\mathbb{P}_0,\ \mathbb{P}_1) \text{ s.t.} \quad & (\mathbb{P}_0,\ \mathbb{P}_1)\in\mathcal{P}, \\ & \forall f\in\mathcal{F},\quad \left|\pi f\left(\mathbb{P}_0\right)+(1-\pi)f\left(\mathbb{P}_1\right)-f\left(\hat{\mathbb{P}}\right)\right|\leq\epsilon_f,\end{aligned} \tag{2}$$

2 Some examples

In this section, we consider a few example instantiations of the general framework described above. For simplicity, we focus on the case where $X_j\in[0,\ 1]$ for all $1\leq j\leq d$. The requirement that $\mathbb{P}_z$ be supported on $[0,\ 1]$ will be implicit throughout our discussion.

We first analyze the case where $\mathcal{P}$ consists of distributions with "poorly separated" means. In particular, let $0<\Delta_j<1$ be user-specified constants and define

$$\mathcal{P}=\mathcal{P}\left(\Delta\right)=\left\{(\mathbb{P}_0,\ \mathbb{P}_1):\ \left|\mathbb{E}_{\mathbb{P}_1}\left[X_j\right]-\mathbb{E}_{\mathbb{P}_0}\left[X_j\right]\right|\leq\Delta_j\right\} \tag{3}$$

A very simple case under this definition of $\mathcal{P}$ arises if we take $\mathcal{F}=\{x\mapsto x_j\}_{j=1}^d$. Indeed, if we let $\hat{\mu}=\frac{1}{n}\sum_{i=1}^n X_i$, then we can write the full set of constraints on $\mu_{0:1}$ as

$$\begin{aligned}0\leq\mu_{z,j}\leq 1,\quad z\in\{0,\ 1\},1\leq j\leq d,\\ -\epsilon_j\leq\pi\mu_{0,j}+(1-\pi)\mu_{1,j}-\hat{\mu}_j\leq\epsilon_j,\quad 1\leq j\leq d,\\ -\Delta_j\leq\mu_{1,j}-\mu_{0,j}\leq\Delta_j,\quad 1\leq j\leq d.\end{aligned}$$

It is easy to see that these joint constraints imply marginal constraints on $\mu_{z,j}$. Indeed, through some simple algebra, we find

$$\max\left\{\hat{\mu}_j-\epsilon_j-(1-\pi)\Delta_j,\ 0\right\}\leq\mu_{0,j}\leq\min\left\{\hat{\mu}_j+\epsilon_j+(1-\pi)\,\Delta_j,\ 1\right\}$$

If some or all of the Δ_j are small, the resulting constraints can be a substantial tightening of the ordinary first-order constraints that arise from the fact that $\mu_{z,j}\in[0,\ 1]$, which have the form

$$\max\left\{\frac{\hat{\mu}_j-\epsilon_j-(1-\pi)}{\pi},\ 0\right\}\leq\mu_{0,j}\leq\min\left\{\frac{\hat{\mu}_j+\epsilon_j}{\pi},\ 1\right\}.$$

If $\mathcal{F}$ just consists of the coordinate mappings as above, we can easily generate the marginal constraints by hand, or even solve the feasibility program exactly using linear programming methods. Provided we are willing to forego these computational advantages, however, we can obtain even tighter constrains by enlarging $\mathcal{F}$ to include $\{x\mapsto x_jx_k\}_{1\leq j<k\leq d}$. Assuming the associated confidence interval widths are ϵ_{jk}, this enlargment adds the constraints

$$-\epsilon_{jk}\leq\pi\mu_{0,j}\mu_{0,k}+(1-\pi)\mu_{1,j}\mu_{1,k}+\rho_{jk}-\hat{S}_{jk}\leq\epsilon_{jk}, \tag{4}$$

where $\hat{S}=\frac{1}{n}\sum_i X_iX_i^T$ is the empirical second moment and $\rho_{jk}=\mathbb{E}\left[\left(X_j-\mu_{Z,j}\right)\left(X_k-\mu_{Z,k}\right)\right]$. Now, we observe that although ρ_{jk} is not directly observable, it does satisfy

$$\begin{aligned}\rho_{jk}=\mathbb{E}\left[\left(X_j-\mu_j\right)\left(X_k-\mu_k\right)\right]+\mathbb{E}\left[\left(\mu_{Z,j}-\mu_j\right)\left(\mu_{Z,k}-\mu_k\right)\right]\\ -\mathbb{E}\left[\left(X_j-\mu_j\right)\left(\mu_{Z,k}-\mu_k\right)\right]-\mathbb{E}\left[\left(\mu_{Z,j}-\mu_j\right)\left(X_k-\mu_k\right)\right].\end{aligned}$$

In particular, we have

$$|\rho_{jk} - \mathbb{E}\left[(X_j - \mu_j)\,(X_k - \mu_k)\right]| \leq \Delta_j\Delta_k + \mathbb{E}\left[|X_j - \mu_j|\right]\Delta_k + \mathbb{E}\left[|X_k - \mu_k|\right]\Delta_j$$

If we suppose $\hat{\lambda}_j$ is a computable upper bound on $\mathbb{E}\left[|X_j - \mu_j|\right]$ and similarly for k and $\hat{\lambda}_k$, then this fact actually allows us to deduce a modified form of the previous constraint that does not involve $\hat{S}_{jk}$ but only $\hat{\mu}_j$ and $\hat{\mu}_k$, viz.

$$-\tilde{\epsilon}_{jk} - \Delta_{jk} \leq \pi\mu_{0,j}\mu_{0,k} + (1-\pi)\mu_{1,j}\mu_{1,k} - \hat{\mu}_j\hat{\mu}_k \leq \tilde{\epsilon}_{jk} + \Delta_{jk}, \tag{5}$$

where $\tilde{\epsilon}_{jk}$ is the width of the confidence interval around $\hat{\mu}_j\hat{\mu}_k$ derived from the widths ϵ_ℓ around $\hat{\mu}_\ell$ for $1 \leq \ell \leq d$ and $\Delta_{jk} = \Delta_j\Delta_k + \hat{\lambda}_j\Delta_k + \hat{\lambda}_k\Delta_j$. In the special case where we know the μ_ℓ exactly (i.e. we do not account for uncertainty in our estimates), the constraint simplifies to

$$-\Delta_{jk} \leq \pi\mu_{0,j}\mu_{0,k} + (1-\pi)\mu_{1,j}\mu_{1,k} - \hat{\mu}_j\hat{\mu}_k \leq \Delta_{jk} \tag{6}$$

Another interesting assumption, which we explore more briefly, consists of *stationarity* of the X_j for $1 \leq j \leq d$. Specifically, we may assume that the joint distribution $(X_j - \mu_{Z,j},\ X_{j+s} - \mu_{Z,j+s})$ is the same for all $1 \leq j \leq d-s$ if the gap $0 < s < d$ is fixed. While this is weaker than the assumption that the X_j form a stationary Markov chain, the stationarity assumption alone is already quite powerful. For instance, it already reduces the number of cross second-moment degrees of freedom to $d-1$ from $\binom{d}{2}$, which in applications can lead to tightening of the first-order partial identification set.

3 Connection to Fréchet-Hoeffding bounds

In this section, we explain how our framework relates to the classical subject of Frchet-Hoeffding (FH) bounds. Here we specialize to the binary case $X_i \in \{0,\ 1\}$, where all questions become equivalent to questions about contingency tables.

One important difference between our proposed methodology and FH bounds is that, ithout independence assumptions, FH bounds provide no information whatever. On the other hand, all of our discussion in the previous section goes through in this case in order to provide FH-type bounds on $\mu_{z,j} = \frac{\mathbb{P}(X_j=1,\ Z=z)}{\mathbb{P}(Z=z)}$ and therefore on $\mathbb{P}\left(X_j = 1,\ Z = z\right)$, which is a cell in the contingency table. Since our feasibility programs (1) and (2) operate at the level of *distributions*, solving either one provides FH-type bounds for free. Indeed, if $\hat{\mathcal{P}} \subset \mathcal{P}$ is the subset of distribution pairs that solve the feasibility program, then for any cell in the contingency table—corresponding to, say, $\mathbb{P}\left(X_{j_1} = b_1, \cdots, X_{j_m} = b_m,\ Z = z\right)$, then we can reach the FH-like conclusion that with probability $\geq 1-\alpha$,

$$\begin{aligned}\inf_{\hat{\mathbb{P}}\in\hat{\mathcal{P}}} \hat{\mathbb{P}}\left(X_{j_1} = b_1, \cdots, X_{j_m} = b_m,\ Z = z\right) &\leq \mathbb{P}\left(X_{j_1} = b_1, \cdots, X_{j_m} = b_m,\ Z = z\right) \\ &\leq \sup_{\hat{\mathbb{P}}\in\hat{\mathcal{P}}} \hat{\mathbb{P}}\left(X_{j_1} = b_1, \cdots, X_{j_m} = b_m,\ Z = z\right)\end{aligned}$$

Since this bound is estimated from data, it comes only with a *probabilistic* guarantee, but the same would be true if FH bounds were stated in terms of probabilities, rather than counts, and if those probabilities were themselves estimated from data (as they would necessarily be in most instances).