Counterfactual Sensitivity and Robustness

Kei Ikegami

September 13, 2023

1 (Generalized) Method of Moments

(Generalized) Method of moments (GMM) is a method to estimate a (semi) parametric model. Below you can find an example in a discrete choice model.

Example 1. *(Discrete Choice) Imagine there is a set of agents: $i \in \{1, \cdot, N\}$. From the researcher's viewpoint, they are all equivalent. They face a choice problem from an alternative set denoted by $\mathcal{J} = \{0, 1, \cdots, J\}$, where 0 denotes choosing the outside option. For example, they face a set of cars and buy just one car or decide not to buy anything. All the alternatives are characterized by a vector $Z_j \in \mathbb{R}^d$. Now the utility attained when an agent chooses j is parametrized by $\theta \in \mathbb{R}^d$ as follows:*

$$u_j = Z_j'\theta + \epsilon_j,$$

where $\epsilon = (\epsilon_0, \cdots, \epsilon_J)$ is an unobserved utility term from the researcher's viewpoint, which follows a known distribution F. ϵ is drawn independently for each agent i, so when necessary I denote it by ϵ^i. By the knowledge of F, we can compute the choice probabilities for each alternative as follows:

$$\begin{aligned} p(j; \theta, F) &= Pr\left(u_j \geq u_{j'} \text{ for all } j' \in \mathcal{J}\right) \\ &= Pr\left(\epsilon_j - \epsilon_{j'} \geq (Z_{j'} - Z_j)'\,\theta \text{ for all } j' \in \mathcal{J}\right). \end{aligned}$$

From data, we can estimate the choice probabilities, $\hat{p}(0), \cdots \hat{p}(J)$, in the following way: where D_i indicates the choice of i,

$$\hat{p}(j) = \frac{\sum_i \mathbf{1}\{D_i = j\}}{N}.$$

GMM searches θ which matches $p(j;\theta,F)$ with $\hat{p}(j)$ as well as possible. In particular, MM searches the parameters satisfying the following condition:

$$\hat{p}(j) = \mathbb{E}^F\left[\mathbf{1}\left\{\epsilon_j - \epsilon_{j'} \geq (Z_{j'} - Z_j)'\,\theta \text{ for all } j' \in \mathcal{J}\right\}\right],\ \forall\ j \in \mathcal{J}.$$

Because the right hand side of the above condition is an expectation, or a moment, this condition is called moment condition. For example, when we assume a type-I extreme value distribution as F, we can write the moment conditions as follows:

$$\hat{p}(j) = \frac{e^{Z_j'\theta}}{\sum_{j' \in \mathcal{J}} e^{Z_{j'}'\theta}},\ \forall\ j \in \mathcal{J}.$$

When the number of moment conditions, now the number of alternatives, is larger than the dimension of the parameter, d, this model is called *over-identified model*. When the other is the case, that is called *under-identified model*. When the two are the same, the model is called *just-identified model*.

Example 1. *(Continued) We denote the vector of the moments and the observed choice probabilities by $m(\theta, F) \in \mathbb{R}^{J+1}$ and $\hat{p} \in \mathbb{R}^{J+1}$. The estimation problem is defined as follows:*

$$\min_{\theta} \ (\hat{p} - m(\theta, F))' \, \Omega \, (\hat{p} - m(\theta, F))$$

where Ω is called a weighting matrix and for now we set $\Omega = I_{J+1}$. In other words, in the most simple case of GMM is the minimization problem of the summation of the squared gap between the model implied moments and the observed moments. This estimator has several nice properties such as consistency and asymptotic normality.

In general notation, we use the following to express the moment conditions:

$$P = \mathbb{E}^F \left[g(\epsilon, \theta) \right].$$

In the above example, $P = \hat{p}$ and $g(\epsilon, \theta) = \left(\mathbf{1} \left\{ \epsilon_j - \epsilon_{j'} \geq (Z_{j'} - Z_j)' \, \theta \text{ for all } j' \in \mathcal{J} \right\} \right)_{J+1}$.

2 Counterfactual Analysis

Economist often tries to simulate the model based on the estimation to know "if" story. This is called counterfactual analysis. In this paper, we focus on one type of such simulation, our target is just one-dimensional object like social welfare.

Example 1. *(Continued) The social welfare is the summation of the achieved utilities in the market. In this example, that is defined as follow: we scaled it by $\frac{1}{N}$*

$$\frac{1}{N} \sum_i u^i_{D_i} = \frac{1}{N} \sum_i Z'_{D_i} \theta + \epsilon^i_{D_i} = \frac{1}{N} \sum_i \max_j \left\{ Z'_j \theta + \epsilon^i_j \right\}.$$

Note that this is random variable because we do not know he realization of ϵ^i for each i. Instead our simulation target is the limit of the social welfare in the infinite number of agent case:

$$W(\theta) = \mathbb{E}^F \left[\max_j \left\{ Z'_j \theta + \epsilon_j \right\} \right].$$

In general notation, we express this object of the counterfactual using $\mathbb{E}^F \left[k(\epsilon, \theta) \right]$. In the above example, $k(\epsilon, \theta) = \max_j \left\{ Z'_j \theta + \epsilon_j \right\}$.

For example, when we use a type I extreme value distribution as F, $W(\theta) = \ln \left(\sum_j e^{Z'_j \theta} \right)$. So after estimating θ using GMM, we plug in it into $W(\theta)$ to get the target value.

3 What this paper does

This paper proposes a sensitivity analysis about the counterfactual analysis. The focus is the sensitivity w.r.t. the distribution of the unobserved terms, F. You know that in the above sections F is used twice: in the estimation process, we derive the moment conditions using the knowledge of F and in the counterfactual analysis, we use the knowledge about F to define the objective value. Hence, if the specification is faw away from the true distribution, the obtained counterfactual values are very different from the true values. The objective of the sensitivity analysis is to provide a range of the target values in the case of the misspecifications of the distribution.

This paper does not take the two steps as in the above sections. Instead, it directly computes the worst case and the best case of the target value within a neighborhood of the specification. Hereafter, we focus on the best case and the same argument can be applied to the worst case. It solves the following problem: given a base line specification $F^\star$ and observed moment values P,

$$(\mathrm{P}_0)\left[\begin{array}{cc} \sup\limits_{\theta\in\Theta, F\in N_\delta(F^\star)} & \mathbb{E}^F\left[k(\epsilon,\theta)\right] \\ \text{s.t.} & P=\mathbb{E}^F\left[g(\epsilon,\theta)\right]. \end{array}\right.$$

Now we define $N_\delta(F^\star)$ as the neighborhood of $F^\star$ with radius $\delta > 0$. The distance between distributions is ϕ-divergence. When we have a convex function $\phi : [0,\infty) \to \mathbb{R}_+$, the definition is as follow:

$$\mathcal{N}_\delta = \left\{F \in \mathcal{F} : D_\phi\left(F \| F_*\right) \le \delta\right\},$$

$$D_\phi\left(F\|F_*\right)=\begin{cases} \int \phi\left(\frac{\mathrm{d}F}{\mathrm{d}F_*}\right)\mathrm{d}F_* & \text{if } F \ll F_*, \\ +\infty & \text{otherwise,} \end{cases}$$

This is a fairly general class of distances between two measures. An example of this is *KL divergence* when $\phi(x) = x\ln x - x + 1$:

$$D_\phi\left(F\|F_*\right)=\int\left(\frac{f}{f_\star}\ln\frac{f}{f_\star}-\frac{f}{f_\star}+1\right)f_\star\mathrm{d}x=\int f\ln\frac{f}{f_\star}-f+f_\star\mathrm{d}x=\int f\ln\frac{f}{f_\star}\mathrm{d}x.$$

Note that $\phi(x) = x\ln x$ gives the same KL divergence, but the function can take negative values.

3.1 Duality

First of all, (P_0) is an infinite-dimensional problem, where we have to optimize w.r.t. F. This is hard to solve. The main benefit of considering the dual problem of (P_0) is that we get a finite-dimensional problem.

We consider to decompose (P_0) into a two step optimization problem:

$$(\mathrm{P}_1)\left[\begin{array}{ccc} \sup\limits_{\theta\in\Theta} & \sup\limits_{F\in N_\delta(F^\star)} & \mathbb{E}^F\left[k(\epsilon,\theta)\right] \\ & \text{s.t.} & P=\mathbb{E}^F\left[g(\epsilon,\theta)\right]. \end{array}\right.$$

And we define the solution of the profiled optimization problem in (P_1) given a value of P as $\bar{K}_\delta(\theta; P, F^\star)$. Then (P_1) is written as follows:

$$\sup_{\theta\in\Theta} \bar{K}_\delta(\theta; P, F^\star).$$

Consider the inner problem:

$$\begin{cases} \sup_F & \mathbb{E}^F\left[k(\epsilon,\theta)\right] \\ \text{s.t.} & \int \phi\left(\frac{\mathrm{d}F}{\mathrm{d}F_*}\right)\mathrm{d}F_* \leq \delta \\ & \int \frac{\mathrm{d}F}{\mathrm{d}F_*}\mathrm{d}F_* = 1 \\ & P = \mathbb{E}^F\left[g(\epsilon,\theta)\right] \end{cases}$$

Now make all the expectation w.r.t. the base line distribution. The last constraint is now $\mathbb{E}^{F_\star}\left[g(\epsilon,\theta)\frac{\mathrm{d}F}{\mathrm{d}F_*}(\epsilon)\right]$. Now we define $m \equiv \frac{\mathrm{d}F}{\mathrm{d}F_*}$. This gives us the following problem:

$$\begin{cases} \sup_F & \mathbb{E}^{F_\star}\left[m(\epsilon)k(\epsilon,\theta)\right] \\ \text{s.t.} & \mathbb{E}^{F_\star}\left[\phi(m(\epsilon))\right] \leq \delta \\ & \mathbb{E}^{F_\star}\left[m(\epsilon)\right] = 1 \\ & P = \mathbb{E}^{F_\star}\left[m(\epsilon)g(\epsilon,\theta)\right] \end{cases}$$

Define the following notations: these are inner products,

$$\begin{cases} < m, k >= \mathbb{E}^{F_\star}\left[m(\epsilon)k(\epsilon,\theta)\right] \\ < m, 1 >= \mathbb{E}^{F_\star}\left[m(\epsilon)\right] \\ < m, g >= \mathbb{E}^{F_\star}\left[m(\epsilon)g(\epsilon,\theta)\right]. \end{cases}$$

And additionally, we write $Q_\phi(m) = \mathbb{E}^{F_\star}\left[\phi(m(\epsilon))\right]$. Then , our inner problem is as follows:

$$(\mathrm{P}_2)\begin{cases} \sup_m & < m, k > \\ \text{s.t.} & Q_\phi(m) \leq \delta \\ & < m, 1 >= 1 \\ & < m, g >= P \end{cases}$$

Add parameters: y_1, y_2 and y_3, and define the parametrized version of the problem including the constraints in objective function.

$$\sup_m \; < m, k > + \mathcal{I}_C\left(Q_\delta(m) - \delta + y_1, < m, 1 > -1 + y_2, < m, g > -P + y_3\right),$$

where

$$\mathcal{I}_C\left(a_1, a_2, a_3\right) = \begin{cases} 0 & \text{if } a_1 \leq 0, a_2 = 0, a_3 = 0 \\ -\infty & \text{otherwise.} \end{cases}$$

The objective function is denoted by $\varphi(m, y)$, which take minus infinity when the control variable, m, is outside of the feasible region. Define the value function given y as

$$(P_y) \qquad v(y) = \sup_m \varphi(m, y) = -\inf_m -\varphi(m, y).$$

The value of the original inner problem is $v(0)$. Under some assumptions, which is denoted by Φ in the paper, we can show the following facts about ϕ and v.

1. ψ is proper and convex,
2. v is proper, convex, and l.s.c. on $\mathbb{R}^{d+2}$,
3. P_y exists for each y.

Consider the dual problem of (P_y). First we take the conjugate of $-v(\cdot)$: where $y^\star$ is the dual variable,

$$\begin{aligned} (-v)^\star(y^\star) &= \sup_y \{y^{\star'} y - -(v(y))\} \\ &= \sup_y \{y^{\star'} y + v(y)\} \\ &= \sup_y \{y^{\star'} y - \inf_m -\varphi(m, y)\} \\ &= \sup_{y,m} \{y^{\star'} y + \varphi(m, y)\}. \end{aligned}$$

Now we define the conjugate function of $-\varphi(m, y)$ as follows:

$$(-\varphi)^\star(m^\star, y^\star) \equiv \sup_{m,y} \{< m^\star, m > + y^{\star'} y + \varphi(m, y)\}.$$

Then, $(-\varphi)^\star(0, y^\star) = \sup_{m,y}\{y^{\star'} y + \varphi(m, y)\}$. Hence we have the following:

$$(-v)^\star(y^\star) = (-\varphi)^\star(0, y^\star).$$

Again we take the conjugate of $(-v)^\star$, where the variable is original y,

$$\begin{aligned} (D_y) \qquad (-v)^{\star\star}(y) &= \sup_{y^\star} \{y^{\star'} y - (-v)^\star(y^\star)\} \\ &= \sup_{y^\star} \{y^{\star'} y - (-\varphi)^\star(0, y^\star)\}. \end{aligned}$$

This problem is called *conjugate dual* of (P_y). When $y = 0$, this corresponds to the *dual* of (P_0): $\sup_{y^\star} -(-\varphi)^\star(0, y^\star)$.

Compute $(-\varphi)^\star(0, y^\star)$ in our case:

$$(-\varphi)^\star(0, y^\star) = \sup_{m,y} \left\{ y^{\star\prime} y + < m, k > + \mathcal{I}_C \left(Q_\delta(m) - \delta + y_1, < m, 1 > -1 + y_2, < m, g > -P + y_3 \right) \right\}.$$

Consider fixing m and optimizing w.r.t. y:

$$\sup_y \left\{ y^{\star\prime} y + < m, k > + \mathcal{I}_C \left(Q_\delta(m) - \delta + y_1, < m, 1 > -1 + y_2, < m, g > -P + y_3 \right) \right\}.$$

Remember that $\mathcal{I}_C$ is negative infinite when it takes the values outside C. So, in the above infimum problem, y is required to make $(Q_\delta(m) - \delta + y_1, < m, 1 > -1 + y_2, < m, g > -P + y_3)$ in C: i.e.

$$Q_\delta(m) - \delta + y_1 \leq 0, \; < m, 1 > -1 + y_2 = 0, \; < m, g > -P + y_3 = 0,$$

Hence, we have.

$$\begin{cases} y_1 \leq -\left(Q_\delta(m) - \delta\right) \\ y_2 = -(< m, 1 > -1) \\ y_3 = -(< m, g > -P). \end{cases}$$

Lastly, we have to decide the value of y_1. Our problem is now

$$\sup_{y_1 \leq -(Q_\delta(m) - \delta)} \{ y_1^\star y_1 - y_2^\star(< m, 1 > -1) - y_3^\star(< m, g > -P) + < m, k > \}.$$

The solution depends on the sign of $y_1^\star$: when $y_1^\star \geq 0$, then the value is maximized when $y_1 = -\left(Q_\delta(m) - \delta\right)$, otherwise the value can be arbitrarily large by taking y_1 small. This leads to the following minimization problem w.r.t. m, which has been fixed,

$$(-\varphi)^\star(0, y^\star) = \sup_m \{ -y_1^\star \left(Q_\delta(m) - \delta\right) - y_2^\star(< m, 1 > -1) - y_3^\star(< m, g > -P) + < m, k > \} + \mathcal{I}_{C^0}(y_1^\star),$$

where $\mathcal{I}_{C^0}(y_1^\star) = \infty$ when $y_1^\star < 0$ and 0 otherwise.

Now we write $y_1^\star = \eta, y_2^\star = \zeta, y_3^\star = \lambda$. And $(\eta, \zeta, \lambda) \in C^0$. Then,

$$\begin{aligned} (-\varphi)^\star(0, (\eta, \zeta, \lambda)) &= \sup_m \left(-\eta Q_\delta(m) - \zeta < m, 1 > -\lambda < m, g > + < m, k > \right) + \eta\delta + \zeta + \lambda' P \\ &= \sup_m \mathbb{E}^{F_\star} \left[m(\epsilon) \left(k(\epsilon) - \zeta - \lambda' g(\epsilon) \right) - \eta \phi(m(\epsilon)) \right] + \eta\delta + \zeta + \lambda' P. \end{aligned}$$

Next, we consider the problem w.r.t. m. Since the class of m is *decomposable*, we can change the order of sup and $\mathbb{E}^{F_\star}$: i.e. we can consider the maximization problem at each point,

$$\mathbb{E}^{F_\star} \left[\sup_m \; m(\epsilon) \left(k(\epsilon) - \zeta - \lambda' g(\epsilon) \right) - \eta \phi(m(\epsilon)) \right].$$

Now $m(\epsilon)$ is just a variable and so the inside of the expectation can be seen as a conjugate function of $\eta\phi(t)$, which is computed as

$$(\eta\phi)^\star(x) \equiv \sup_t\{tx - \eta\phi(t)\} = \begin{cases} \eta\phi^\star(\frac{x}{\eta}) & \text{if } \eta > 0 \\ 0 & \text{if } \eta = 0 \text{ and } x \le 0 \\ +\infty & \text{if } \eta = 0 \text{ and } x > 0. \end{cases}$$

This gives

$$(-\varphi)^\star(0,(\eta,\zeta,\lambda)) = \eta\mathbb{E}^{F_\star}\left[\phi^\star\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\right] + \eta\delta + \zeta + \lambda' P.$$

By strong duality, we know that $-v(0) = (-v)^{\star\star}(0)$. And from (D_y), we have

$$\begin{aligned}
-v(0) = (-v)^{\star\star}(0) &= \sup_{\eta,\zeta,\lambda}\{-(-\varphi)^\star(0,(\eta,\zeta,\lambda))\} \\
&= \sup_{\eta,\zeta,\lambda} -\left(\eta\mathbb{E}^{F_\star}\left[\phi^\star\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\right] + \eta\delta + \zeta + \lambda' P\right).
\end{aligned}$$

This gives:

$$\begin{aligned}
v(0) &= -\sup_{\eta,\zeta,\lambda} -\left(\eta\mathbb{E}^{F_\star}\left[\phi^\star\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\right] + \eta\delta + \zeta + \lambda' P\right) \\
&= \inf_{\eta,\zeta,\lambda} \eta\mathbb{E}^{F_\star}\left[\phi^\star\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\right] + \eta\delta + \zeta + \lambda' P. \qquad (1)
\end{aligned}$$

This corresponds to (14) in Christensen and Connault (2023).

Example 2. *When we use KL divergence, we have $\phi^\star(x) = e^x - 1$. Then,*

$$\inf_{\eta>0,\zeta\in\mathbb{R},\lambda\in\mathbb{R}} \eta\mathbb{E}^{F_\star}\left[e^{\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}} - 1\right] + \eta\delta + \zeta + \lambda' P.$$

Now we solve w.r.t. ζ. By considering FOC. we get the following:

$$e^{-\frac{\zeta}{\eta}}\mathbb{E}^{F_\star}\left[e^{\frac{k(\epsilon)-\lambda' g(\epsilon)}{\eta}}\right] = 1.$$

By inserting this, we get the following problem w.r.t. η and λ:

$$\bar{K}_\delta(\theta; P, F_\star) = \inf_{\eta>0,\lambda\in\Lambda} \eta\log\mathbb{E}^{F_*}\left[e^{(k(\epsilon,\theta)-\lambda' g(\epsilon,\theta))/\eta}\right] + \eta\delta + \lambda' P.$$

Finally, the upper bounds of the counterfactual value is obtained by the following problem:

$$\sup_{\theta\in\Theta}\inf_{\eta>0,\lambda\in\Lambda} \eta\log\mathbb{E}^{F_*}\left[e^{(k(\epsilon,\theta)-\lambda' g(\epsilon,\theta))/\eta}\right] + \eta\delta + \lambda' P.$$

3.2 Least/Best Favorable Distribution

Consider the FOC of problem (1). The assumption Φ in the main text assures the FOC is sufficient for the optimal solution.

Take FOC w.r.t. ζ:

$$\eta\mathbb{E}^{F_\star}\left[-\frac{1}{\eta}\phi^{\star\prime}\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\right]+1=-\mathbb{E}^{F_\star}\left[\phi^{\star\prime}\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\right]+1=0.$$

This gives $\mathbb{E}^{F_\star}\left[\phi^{\star\prime}\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\right]=1$, which implies the following:

$$\int\phi^{\star\prime}\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)f^\star(\epsilon)\mathrm{d}\epsilon=1.$$

So we can consider $\phi^{\star\prime}\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)$ as a change of measure, $\bar{m}^{\delta,\theta}(\epsilon)$. And using this we define a best favorable distribution $d\bar{F}^{\delta,\theta}=\bar{m}^{\delta,\theta}(\epsilon)dF_\star$.

Assume that $\eta>0$ at the solution. Take FOC w.r.t. η:

$$\begin{aligned}
&\mathbb{E}^{F_\star}\left[\phi^{\star}\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\right]+\eta\mathbb{E}^{F_\star}\left[\phi^{\star\prime}\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\left(-\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta^2}\right)\right]\\
&=\mathbb{E}^{F_\star}\left[\phi^{\star}\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\right]-\mathbb{E}^{F_\star}\left[\phi^{\star\prime}\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\right]+\delta=0\\
&\Rightarrow\ \mathbb{E}^{F_\star}\left[\phi^{\star\prime}\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)-\phi^{\star}\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\right]=\delta.
\end{aligned}$$

Now when we write $x=\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}$, the inside of the expectation is

$$\phi^{\star\prime}(x)x-\phi^\star(x).$$

In the following, we show the following equality:

$$\phi\left(\phi^{\star\prime}(x)\right)=\phi^{\star\prime}(x)x-\phi^\star(x). \tag{2}$$

Proof. We consider the Legendre transform of $\phi^\star(x)$:

$$\phi^{\star\star}(x^\star)\equiv\sup_x\{xx^\star-\phi^\star(x)\}.$$

By Assumption Φ, FOC is enough to solve this and we get the maximizer as $\phi^{\star\prime-1}(x^\star)$. Then we get the following:

$$\phi^{\star\star}(x^\star)=\phi^{\star\prime-1}(x^\star)x^\star-\phi^\star\left(\phi^{\star\prime-1}(x^\star)\right).$$

Now we insert $\phi^{\star\prime}(x)$ into the place of $x^\star$ to get the following:

$$\phi^{\star\star}\left(\phi^{\star\prime}(x)\right)=x\phi^{\star\prime}(x)-\phi^\star\left(\phi^{\star\prime-1}(\phi^{\star\prime}(x))\right)=x\phi^{\star\prime}(x)-\phi^\star(x).$$

By bi-conjugacy, we know that $\phi=\phi^{\star\star}$, then we have the claim:

$$\phi\left(\phi^{\star\prime}(x)\right)=x\phi^{\star\prime}(x)-\phi^\star(x).$$

□

From the above argument, we can write the FOC w.r.t. η as follows:

$$\delta = \mathbb{E}^{F_\star}\left[\phi\left(\phi^{\star\prime}\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)\right)\right] = \mathbb{E}^{F_\star}\left[\phi\left(\bar{m}^{\delta,\theta}(\epsilon)\right)\right] = D_\phi\left(\bar{F}^{\delta,\theta}\|F_*\right).$$

This implies that $\bar{F}^{\delta,\theta} \in N_\delta$.

Take FOC w.r.t. λ:

$$\eta\mathbb{E}^{F_\star}\left[-\frac{1}{\eta}\phi^{\star\prime}\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)g(\epsilon)\right]+P = -\mathbb{E}^{F_\star}\left[\phi^{\star\prime}\left(\frac{k(\epsilon)-\zeta-\lambda' g(\epsilon)}{\eta}\right)g(\epsilon)\right]+P = 0.$$

This implies that $<\bar{m}^{\delta,\theta}, g>= P$, i.e., $\bar{F}^{\delta,\theta}$ satisfies the moment conditions.

Now we have shown that $\bar{m}^{\delta,\theta}$ satisfies the constraints in (P_2). Next, we show that $\bar{m}^{\delta,\theta}$ maximizes the objective function. This is done by showing that the value of the primal problem is equal to the value of the dual problem because we have strong duality result. First, the value of the primal problem is

$$\mathbb{E}^{\bar{F}^{\delta,\theta}}\left[k(\epsilon)\right] = \mathbb{E}^{F_\star}\left[\bar{m}^{\delta,\theta}(\epsilon)k(\epsilon)\right].$$

The value of the dual problem is as follows: let $(\bar{\eta}, \bar{\zeta}, \bar{\lambda})$ be the solution of the dual problem.

$$\begin{aligned}
&\bar{\eta}\mathbb{E}^{F_\star}\left[\phi^\star\left(\frac{k(\epsilon)-\bar{\zeta}-\bar{\lambda}' g(\epsilon)}{\bar{\eta}}\right)\right]+\bar{\eta}\delta+\bar{\zeta}+\bar{\lambda}'P \\
&= \bar{\eta}\left(\mathbb{E}^{F_\star}\left[\bar{m}^{\delta,\theta}(\epsilon)\left(\frac{k(\epsilon)-\bar{\zeta}-\bar{\lambda}' g(\epsilon)}{\bar{\eta}}\right)\right]-\delta\right)+\bar{\eta}\delta+\bar{\zeta}+\bar{\lambda}'P \quad \because \text{FOC w.r.t. } \eta \\
&= \mathbb{E}^{F_\star}\left[\bar{m}^{\delta,\theta}(\epsilon)k(\epsilon)\right]-\mathbb{E}^{F_\star}\left[\bar{m}^{\delta,\theta}(\epsilon)\right]\bar{\zeta}-\mathbb{E}^{F_\star}\left[\bar{m}^{\delta,\theta}(\epsilon)\bar{\lambda}' g(\epsilon)\right]+\bar{\zeta}+\bar{\lambda}'P \\
&= \mathbb{E}^{F_\star}\left[\bar{m}^{\delta,\theta}(\epsilon)k(\epsilon)\right] \quad \because \text{FOCs w.r.t. } \zeta \text{ and } \lambda.
\end{aligned}$$

Hence we conclude that $\bar{F}^{\delta,\theta}$ solves the primal problem.

Example 2. *(continued) When we use KL-divergence, we know that $\phi^{\star\prime}(x) = e^x$. Then, the change of measure to the best favorable distribution is as follows:*

$$\bar{m}^{\delta,\theta}(\epsilon) = \frac{e^{\frac{k(\epsilon)-\bar{\lambda}' g(\epsilon)}{\bar{\eta}}}}{\mathbb{E}^{F_\star}\left[e^{\frac{k(\epsilon')-\bar{\lambda}' g(\epsilon')}{\bar{\eta}}}\right]},$$

where the denominator is an adjustment term.