Combining Datasets

Andrew Goldstein

February 21, 2019

Introduction

The setting we are considering here involves two distinct datasets.

1. Observational data $Z_t = \begin{bmatrix} Y_t \\ X_t \end{bmatrix} \overset{iid}{\sim} P_{\theta^*}, t = 1, \ldots, n_o$;
2. Simulation data $\tilde{Z}_t = \begin{bmatrix} \tilde{Y}_t \\ \tilde{X}_t \end{bmatrix} \overset{iid}{\sim} P_{\tilde{\theta}}, t = 1, \ldots, n_s$.

Our goals are to estimate θ^* and use our estimate and a new X_t to predict the corresponding Y_t. Here, we assume $\tilde{\theta} \neq \theta^*$, so it is unclear if we should use the simulation data in our estimation of θ^*.

This memo considers two alternative for utilizing the simulation data $\tilde{Z}_t$:

1. Treat the Z_t and $\tilde{Z}_t$ as being $\overset{iid}{\sim} P_{\theta^*}$;
2. Use the $\tilde{Z}_t$ to estimate $\tilde{\theta}$, use this to generate a prior for θ^*, then proceed in the Bayesian way to estimate θ^*.

In the following sections, we make the following simplifying assumption:

$$\begin{aligned} P_\theta &= \mathcal{N}(\theta, \Sigma) \\ \Sigma = \begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{YX}^T & \Sigma_{XX} \end{bmatrix} \\ \theta &= \begin{bmatrix} \theta_Y \\ \theta_X \end{bmatrix} \end{aligned}$$

for a known and fixed Σ.

We show that option 1 can be thought of as a Bayesian procedure in which we use the simulation data to form a particular (sensible) prior, thus we can analyze when these procedure outperform ignoring the simulation data under the same framework.

Preliminaries

Before we proceed, we recall the following results relating to the multivariate normal (using the block-matrix representation for Σ, θ:

1. $$Y|X \sim \mathcal{N}\Big(\theta_Y + \Sigma_{YX}\Sigma_{XX}^{-1}(X - \theta_X), \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{YX}^T\Big)$$
2. $$\begin{aligned} X|\theta, \Sigma \sim \mathcal{N}(\mu, \Sigma), \quad \theta \sim \mathcal{N}(\theta_0, \Sigma_0) \Rightarrow \theta|X_1, \ldots, X_n \sim \mathcal{N}(\theta_n, \Sigma_n) \\ \theta_n = \Sigma_0(\Sigma_0 + \frac{1}{n}\Sigma)^{-1}\bar{X} + \frac{1}{n}\Sigma(\Sigma_0 + \frac{1}{n}\Sigma)^{-1})\theta_0 = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1}(\Sigma_0^{-1}\theta_0 + n\Sigma^{-1}\bar{X}) \\ \Sigma_n = \Sigma_0(\Sigma_0 + \frac{1}{n}\Sigma)^{-1}\frac{1}{n}\Sigma = (\Sigma_0^{-1} + n\Sigma^{-1})^{-1} \end{aligned}$$

Option 1: Combine Datasets and Treat as iid

In this section, we consider the option of naively combining all of the data. We proceed as follows:

1. Treat all data as being iid from the same distribution;
2. Estimate $\hat{\theta}^{MLE} = \begin{bmatrix} \hat{\theta}_Y^{MLE} \\ \hat{\theta}_X^{MLE} \end{bmatrix}$;
3. For new observational data, predict $Y|X$ as $\hat{Y}(X) = \mathbb{E}[Y|X]$;
4. Calculate the mean-square prediction error, $\mathbb{E}_{(Y,X)\sim P_{\theta^*}}[\|\hat{Y}(X) - Y\|^2]$.

It is easy to show that the following are the distributions of the MLE estimates:

$$\hat{\theta}^{MLE} \sim \mathcal{N}\left(\frac{n_o\theta^* + n_s\tilde{\theta}}{n_o + n_s}, \frac{1}{n_o + n_s}\Sigma\right)$$
$$\hat{\theta}_Y^{MLE} \sim \mathcal{N}\left(\frac{n_o\theta_Y^* + n_s\tilde{\theta}_Y}{n_o + n_s}, \frac{1}{n_o + n_s}\Sigma_{YY}\right)$$
$$\hat{\theta}_X^{MLE} \sim \mathcal{N}\left(\frac{n_o\theta_X^* + n_s\tilde{\theta}_X}{n_o + n_s}, \frac{1}{n_o + n_s}\Sigma_{XX}\right)$$

Using the expression for the conditional expectation of $Y|X$, we get:

$$\hat{Y}(X) = \begin{bmatrix} I & -\Sigma_{YX}\Sigma_{XX}^{-1} \end{bmatrix} \hat{\theta}^{MLE} + \Sigma_{YX}\Sigma_{XX}^{-1}X$$

Focusing on the first part:

$$\begin{bmatrix} I & -\Sigma_{YX}\Sigma_{XX}^{-1} \end{bmatrix} \hat{\theta}^{MLE} \sim$$
$$\mathcal{N}\left(\frac{n_o\theta_Y^* + n_s\tilde{\theta}_Y}{n_o + n_s} - \Sigma_{YX}\Sigma_{XX}^{-1}\frac{n_o\theta_X^* + n_s\tilde{\theta}_X}{n_o + n_s}, \frac{1}{n_o + n_s}\begin{bmatrix} I & -\Sigma_{YX}\Sigma_{XX}^{-1} \end{bmatrix}\begin{bmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{YX}^T & \Sigma_{XX} \end{bmatrix}\begin{bmatrix} I \\ -\Sigma_{YX}\Sigma_{XX}^{-1} \end{bmatrix}\right) \sim$$
$$\mathcal{N}\left(\frac{n_o\theta_Y^* + n_s\tilde{\theta}_Y}{n_o + n_s} - \Sigma_{YX}\Sigma_{XX}^{-1}\frac{n_o\theta_X^* + n_s\tilde{\theta}_X}{n_o + n_s}, \frac{1}{n_o + n_s}\left[\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{YX}^T\right]\right)$$

Thus, we get the following:

$$\hat{Y}(X) - Y = \begin{bmatrix} I & -\Sigma_{YX}\Sigma_{XX}^{-1} \end{bmatrix}(\hat{\theta}^{MLE} - \begin{bmatrix} Y \\ X \end{bmatrix}) \sim$$
$$\mathcal{N}\left(\frac{n_o\theta_Y^* + n_s\tilde{\theta}_Y}{n_o + n_s} - \Sigma_{YX}\Sigma_{XX}^{-1}\frac{n_o\theta_X^* + n_s\tilde{\theta}_X}{n_o + n_s}, \frac{1}{n_o + n_s}\left[\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{YX}^T\right]\right) -$$
$$\mathcal{N}\left(\theta_Y^* - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_X^*, \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{YX}^T\right) \sim$$
$$\mathcal{N}\left(\frac{n_s}{n_o + n_s}\left[(\tilde{\theta}_Y - \theta_Y^*) - \Sigma_{YX}\Sigma_{XX}^{-1}(\tilde{\theta}_X - \theta_X^*)\right], (1 + \frac{1}{n_o + n_s}(\Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{YX}^T)\right) \sim$$
$$\mathcal{N}\left(\frac{n_s}{n_o + n_s}\theta_Y', (1 + \frac{1}{n_o + n_s})\Sigma_{Y|X}\right)$$

Where $\theta_Y' = \left[(\tilde{\theta}_Y - \theta_Y^*) - \Sigma_{YX}\Sigma_{XX}^{-1}(\tilde{\theta}_X - \theta_X^*)\right] = \mathbb{E}[\tilde{Y}|\tilde{X}] - \mathbb{E}[Y^*|X^*]$, the diffference in conditional expectations between the two distributions. And $\Sigma_{Y|X}$ is the covariance matrix of $Y|X$ under the shared covariance structure between the two distributions.

Let the eigenvalue decomposition of $\Sigma_{Y|X} = Q\Lambda Q^T \in \mathbb{R}^{d\times d}$. Then

$$\hat{Y}(X) - Y \overset{\mathcal{D}}{=} \mathcal{N}\Big(\frac{n_s}{n_o+n_s}\theta'_Y, (1+\frac{1}{n_o+n_s})Q\Lambda Q^T\Big) \overset{\mathcal{D}}{=}$$

$$\sqrt{\frac{n_o+n_s+1}{n_o+n_s}}\mathcal{N}\Big(\frac{n_s}{\sqrt{(n_o+n_s)(n_o+n_s+1)}}\theta'_Y, Q\Lambda Q^T\Big) \overset{\mathcal{D}}{=}$$

$$\sqrt{\frac{n_o+n_s+1}{n_o+n_s}}Q\cdot\mathcal{N}\Big(\frac{n_s}{\sqrt{(n_o+n_s)(n_o+n_s+1)}}Q^T\theta'_Y, \Lambda\Big) \Rightarrow$$

$$\|\hat{Y}(X) - Y\|^2 \overset{\mathcal{D}}{=} \Big(\frac{n_o+n_s+1}{n_o+n_s}\Big)\Big\|\mathcal{N}\Big(\frac{n_s}{\sqrt{(n_o+n_s)(n_o+n_s+1)}}Q^T\theta'_Y, \Lambda\Big)\Big\| \Rightarrow$$

$$\mathbb{E}[\|\hat{Y}(X) - Y\|^2] = \Big(\frac{n_o+n_s+1}{n_o+n_s}\Big)\sum_{i=1}^{d}\lambda_i^2\mathbb{E}\Big[\chi_1^2\Big(\frac{n_s^2}{\lambda_i^2(n_s+n_s)(n_o+n_s+1)}\langle q_i, \theta'_Y\rangle^2\Big)\Big] =$$

$$\Big(\frac{n_o+n_s+1}{n_o+n_s}\Big)\sum_{i=1}^{d}\lambda_i^2\cdot\Big(1+\frac{n_s^2}{\lambda_i^2(n_s+n_s)(n_o+n_s+1)}\langle q_i, \theta'_Y\rangle^2\Big) =$$

$$\Big(\frac{n_o+n_s+1}{n_o+n_s}\Big)\sum_{i=1}^{d}\lambda_i^2 + \Big(\frac{n_o+n_s+1}{n_o+n_s}\Big)\Big(\frac{n_s^2}{(n_s+n_s)(n_o+n_s+1)}\Big)\|Q^T\theta'_Y\|^2 =$$

$$\Big(\frac{n_o+n_s+1}{n_o+n_s}\Big)\sum_{i=1}^{d}\lambda_i^2 + \frac{n_s^2}{(n_o+n_s)^2}\|\theta'_Y\|^2$$

$$\therefore \mathbb{E}_{(Y,X)\sim P_{\theta^*}}[\|\hat{Y}(X) - Y\|^2] = \Big(\frac{n_o+n_s+1}{n_o+n_s}\Big)\sum_{i=1}^{d}\lambda_i^2 + \frac{n_s^2}{(n_o+n_s)^2}\|\theta'_Y\|^2$$

The alternative is that we can ignore the simulation data, i.e. $n_s := 0$, and still look at the conditional expectation. Call this estimator $Y^\dagger(X)$. Then we get

$$\mathbb{E}_{(Y,X)\sim P_{\theta^*}}[\|Y^\dagger(X) - Y\|^2] = \Big(\frac{n_o+1}{n_o}\Big)\sum_{i=1}^{d}\lambda_i^2$$

We can now find conditions under which including the simulation data yields lower mean-square predictive error:

$$\mathbb{E}_{(Y,X)\sim P_{\theta^*}}[\|\hat{Y}(X) - Y\|^2] \le \mathbb{E}_{(Y,X)\sim P_{\theta^*}}[\|Y^\dagger(X) - Y\|^2] \iff$$

$$\Big(\frac{n_o+n_s+1}{n_o+n_s}\Big)\sum_{i=1}^{d}\lambda_i^2 + \frac{n_s^2}{(n_o+n_s)^2}\|\theta'_Y\|^2 \le \Big(\frac{n_o+1}{n_o}\Big)\sum_{i=1}^{d}\lambda_i^2 \iff$$

$$\frac{n_s^2\|\theta'_Y\|^2}{(n_o+n_s)^2} \le \frac{(n_o+1)(n_o+n_s) - n_s(n_o+n_s+1)}{n_o(n_o+n_s)}\sum_{i=1}^{d}\lambda_i^2 \iff$$

$$\frac{n_s^2\|\theta'_Y\|^2}{n_o+n_s} \le \frac{n_o^2+n_on_s+n_o+n_s-n_o^2-n_on_s-n_o}{n_o}\sum_{i=1}^{d}\lambda_i^2 \iff$$

$$\frac{n_s^2\|\theta'_Y\|^2}{n_o+n_s} \le \frac{n_s}{n_s}\sum_{i=1}^{d}\lambda_i^2 \iff$$

$$\frac{n_o\cdot n_s\cdot\|\theta'_Y\|^2}{n_o+n_s} \le \sum_{i=1}^{d}\lambda_i^2$$

Option 2: Use the Simulation Data to Form a Prior for θ^*

In this section, we consider the option of using the simulation data to form a prior for θ^*. We proceed as follows:

1. Use the simulation data to form some prior $\pi(\theta^*)$;
2. Estimate $\hat{\theta} = \mathbb{E}[\theta^*|Z]$;
3. For new observational data, predict $Y|X$ as $\hat{Y}(X) = \mathbb{E}[Y|X]$;
4. Calculate the mean-square prediction error, $\mathbb{E}_{(Y,X)\sim P_{\theta^*}}[\|\hat{Y}(X) - Y\|^2]$.

Option 1 as a Bayesian Procedure

First, we show that option 1 above can be thought of as a Bayesian procedure with a particular prior.

Let $\hat{\tilde{\theta}}^{MLE}$ be the MLE of $\tilde{\theta}$ from the simulation data, $\tilde{Z}_t$. We know that $\hat{\tilde{\theta}}^{MLE} \sim \mathcal{N}(\tilde{\theta}, \frac{1}{n_s}\Sigma)$. From this, use the following prior for θ^*:

$$\theta^* \sim \mathcal{N}\Big(\hat{\tilde{\theta}}^{MLE}, \frac{1}{n_s}\Sigma\Big)$$

Thus, using the preliminary results of the posterior distribution of the Gaussian mean, we get:

$$\theta^*|Z_1, \ldots, Z_{n_o} \sim \mathcal{N}\Big(\frac{n_o\hat{\theta}^{MLE} + n_s\hat{\tilde{\theta}}^{MLE}}{n_o + n_s}, \frac{1}{n_o + n_s}\Sigma\Big)$$

The conditional mean is the same as before, thus the analysis proceeds identically as above.