Final Exam

June 28, 2020
Sourav Sahoo, EE17B040

Problem

You own a construction company which specializes in residential buildings, i.e., large apartments, medium size apartments, bungalows etc. You have to decide at the start the year what your current focus will be for the next six months, i.e. what projects your will undertake based on various factors such as profits in the last year in the various type of projects you undertook, market sentiment, the labour cost and housing available, price of raw materials, the external funding you will be able to get for these projects, buyer sentiment, current liquidity you have etc. Think like a Bayesian and assume priors for all these factors and then come up with a plan on how many units you will think of building and the type of housing you will think of building. Whenever you have a doubt about the way forward make logical assumptions, state them and proceed forward. You can add more factors as long as you can justify them.

Now suddenly there is an unexpected shutdown, people are going to lose jobs, labour may not be available, there are multiple restrictions on a number of people you can employ in one project simultaneously, banks are wary of lending to both buyer and seller. Is all your prior knowledge based on which you came up with a plan now not useful? What will you do now to try to salvage the situation? What about the half completed projects? What will be your strategy now?

Solution

Let there be n dependent factors and m types of buildings that can be built. Let $\mathbf{p} = [p_1, \ldots, p_m]$ indicate the distribution of the buildings where $0 \leqslant p_i \leqslant 1$ for $i \in [m]$ and $\sum_{i=1}^{m} p_i = 1$. Along with $\mathbf{p}$, the total number of buildings that need to be built also needs to be decided. Let that be denoted by k. So, a total of $(m-1)+1 = m$ independent variables need to be estimated. The summation condition on p_i's ensure there are $m-1$ independent values in $\mathbf{p}$. Let $\mathbf{x} \in \mathbb{R}^n$ be the dependent factors, and $\mathbf{y} \in \mathbb{R}^m$ be the variables that have to be estimated.

Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_T]$ be the observed values of dependent vectors for the past T years, and $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_T]$ be the values of the variables that have to estimated for the same period.

An oversimplified case

Let us first try to solve an oversimplified version where $n = 1$ and $m = 1$ i.e., we have to estimate the number of buildings that can be built based on a single factor (say, availability of funds). Suppose the dependent factor $x \sim \mathcal{N}(\theta, \sigma^2)$, where σ is known and θ is unknown. Suppose the prior distribution of θ is given by an oracle as $\mathcal{N}(\theta_0, \sigma_0^2)$. So by Bayesian estimation, we have:

$$p(\theta|x) \propto p(x|\theta)p(\theta) \propto exp\left(-\frac{(x-\theta)^2}{\sigma^2}\right) exp\left(-\frac{(\theta-\theta_0)^2}{\sigma_0^2}\right) = exp\left(-\frac{(\theta-\hat{\theta})^2}{\hat{\sigma}^2}\right)$$

where the values of $\hat{\theta}$ and $\hat{\sigma}$ is given in (1).

$$\begin{aligned} \hat{\sigma}^{-2} &= \sigma^{-2} + \sigma_0^{-2} \\ \hat{\theta} &= \frac{x\sigma_0^2 + \theta_0\sigma^2}{\sigma^2 + \sigma_0^2} \end{aligned} \tag{1}$$

Now, we need to establish a relation between y, the random variable that has to be estimated and θ. Let us introduce a dummy random variable $u \sim \mathcal{N}(\theta, \gamma^2)$ and a transformation ψ such that $y = \psi(u)$. The need for the transformation would be clear in the next subsection. For sake of simplicity let us assume that ψ is a known *linear* transformation and γ is known. Using law of total probability, we have the expression as given in (2).

$$.p(u|x) = \int p(u|\theta, x)p(\theta|x)d\theta = \int p(u|\theta)p(\theta|x)d\theta \tag{2}$$

The last equality holds because u given θ does not depend on x. From (1) and (2), we have the $u|\theta \sim \mathcal{N}(\theta, \gamma^2)$ and $\theta|x \sim \mathcal{N}(\hat{\theta}, \hat{\sigma}^2)$. So, $u|x \sim \mathcal{N}(\hat{\theta}, \gamma^2 + \hat{\sigma}^2)$[1]. Once the distribution of $u|x$ is known, the distribution of $y|x$ can be found as $y = \psi(u) = \Psi u$, for some constant $\Psi \in \mathbb{R}$. So, $y|x \sim \mathcal{N}(\Psi\hat{\theta}, \Psi^2(\gamma^2 + \hat{\sigma}^2))$. We can get a point prediction by an appropriate measure of central tendency. In this case, *mode* is an appropriate measure as it resembles the *maximum aposteriori* (MAP) estimate. Hence, $\boxed{\hat{y} = \Psi\hat{\theta}}$, i.e., the mode of the $p(y|x)$.

The multivariate case

Now, we will consider the case when $n, m > 1$. If we *assume that the dependent factors are independent*, then the case we discussed in the previous section extends here naturally. We have to find the distribution of $\boldsymbol{\theta} = [\theta_1, \ldots, \theta_n]$ and distribution of each $\theta_i, i \in [n]$ can be found by the method discussed earlier i.e.,

$$p(\theta_i|x_i) \propto p(x_i|\theta_i)p(\theta_i) = exp\left(-\frac{(\theta_i - \hat{\theta}_i)^2}{\hat{\sigma_i}^2}\right)$$

with $\hat{\theta}_i$ and $\hat{\sigma}_i^2$ being similar to the ones calculated in (1). Let $\hat{\boldsymbol{\theta}} = [\hat{\theta}_1, \ldots, \hat{\theta}_n]$ and $\widehat{\boldsymbol{\Sigma}}$ is a diagonal matrix with $\widehat{\boldsymbol{\Sigma}}_{ii} = \hat{\sigma}_i^2$ for $i \in [n]$.

If we assume the factors are not independent, then assume $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Sigma})$ with $\boldsymbol{\Sigma}$ known and the prior of $\boldsymbol{\theta}$ be $\mathcal{N}(\boldsymbol{\theta}_0, \boldsymbol{\Sigma}_0)$. Now, if we calculate the posterior distribution, we get a normal distribution with parameters $\hat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\Sigma}}$ with values as given in (3).

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}^{-1} &= \boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1} \\ \hat{\boldsymbol{\theta}} &= \left(\boldsymbol{\Sigma}^{-1} + \boldsymbol{\Sigma}_0^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\theta}_0 + \boldsymbol{\Sigma}^{-1}\mathbf{x}\right) \end{aligned} \tag{3}$$

Instead of a scalar, the dummy random variable here is $\mathbf{u} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Gamma})$ and $\mathbf{y} = \psi(\mathbf{u}) = \Psi\mathbf{u}$ where $\Psi \in \mathbb{R}^{m \times n}$. *The transformation matrix Ψ maps the n-dimensional dependent factors to the m-dimensional space in which the variables needs to be estimated.* Extending the equation in (2) for vectors and using $\mathbf{u}|\boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{\theta}, \boldsymbol{\Gamma})$ and $\boldsymbol{\theta}|\mathbf{x} \sim \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\Sigma}}\right)$, we get $\mathbf{u}|\mathbf{x} \sim \mathcal{N}\left(\hat{\boldsymbol{\theta}}, \boldsymbol{\Gamma} + \widehat{\boldsymbol{\Sigma}}\right)$. As $\mathbf{y} = \Psi\mathbf{u}$, we have, $\mathbf{y}|\mathbf{x} \sim \mathcal{N}\left(\Psi\hat{\boldsymbol{\theta}}, \Psi\left(\boldsymbol{\Gamma} + \widehat{\boldsymbol{\Sigma}}\right)\Psi^\top\right)$. Hence, $\boxed{\hat{\mathbf{y}} = \Psi\hat{\boldsymbol{\theta}}}$.

Unexpected shutdown

In case of an unexpected shutdown, we need to analyse *uncertainty* more carefully. It refers to uncertainty in the number of workers available, uncertainty in people's purchasing power, and other such relevant dependent factors. In the scenario, as mentioned earlier, this uncertainty is represented by the *variance of the normal distribution*. So, in such a scenario we assume that now a dependent factor $x \sim \mathcal{N}(\theta, \sigma'^2)$ where $\sigma' > \sigma$. So, from (1), we get that $\hat{\sigma}' > \hat{\sigma}$, i.e., the new posterior variance will be higher than the old posterior variance. In the last step, when we derive the distribution of $y|x$, we also find that the new

[1] https://stats.stackexchange.com/questions/204317/

distribution has a much *higher variance* as compared to the previous case. This observation reflects that in the new scenario, the *uncertainty in the estimated values are higher* than the previous case, which also implies that now the *risk involved has increased.* Although it should be noted that as we have *assumed* a normal distribution, the mode is independent of the change in variance. However, if we had chosen some other distribution such as Rayleigh i.e., $p(x;\sigma) = \frac{x}{\sigma^2}e^{-x^2/2\sigma^2}$, then changing the variance also changes the mode.

As the risk involved in the business increases due to the shutdown, risk management steps need to be taken. For example:

- More resources should be diverted towards the constructions with a lower risk involved, such as medium-sized buildings in and near the city, as they would be more affordable and desirable to people than bungalows post-shutdown.
- All the high risk and ambitious projects should be halted or provided minimal resources until the situation becomes normal.
- Adequate resources must also be utilized for the protection and well being of the employees so that the company can be bounce back to regular business as soon as possible.

Refining the solution

In the above subsections, we have assumed a lot of the statements to make our calculations easier and keep the integrals tractable. In this subsection, we clarify some of the details:

- As most of the factors and the estimated variables are positive valued in the real-life scenario, it is preferable to choose a distribution that has a support $(0, \infty)$ such as the Rayleigh distribution or the Gamma distribution. A different approach could be as follows: instead of using the raw values of the different factors involved, they could be normalized, i.e., by subtracting the sample mean and dividing by the sample variance.
- We have assumed the mean and variance of the prior distribution. Hierarchical modeling can help in removing the bias that is caused by our assumption of a particular prior.
- We have also assumed that the transformation matrix Ψ is known. Suppose, the value is unknown, an estimate of Ψ can be done in the following method. Suppose, $\sigma^2/\sigma_0^2 \ll 1$, then $\hat{\theta} \approx x$ from (1). So, $\hat{y} \approx \Psi x$. Now, according to our assumption we have historical data for the past T years which implies $\mathbf{Y} = \Psi\mathbf{X}$. By ordinary least squares, $\widehat{\Psi} = \mathbf{Y}\mathbf{X}^\top\left(\mathbf{X}\mathbf{X}^\top\right)^{-1}$.
- If we want to further incorporate weights to the observations based on the profits of that particular year and age of the observation, we can also do a weighted least squares, which would result in $\widehat{\Psi'} = \mathbf{Y}\mathbf{\Omega}^{-1}\mathbf{X}^\top\left(\mathbf{X}\mathbf{\Omega}^{-1}\mathbf{X}^\top\right)^{-1}$ where $\mathbf{\Omega} \in \mathbb{R}^{T\times T}$ is a diagonal matrix with $\mathbf{\Omega}_{tt} = w_t$ for $t \in [T]$. An appropriate choice for w_t could be $e^{-t}/(1 + e^{-p_t})$ where p_t is the net profit for the year t. As p_t increases, w_t increases for a constant t (more weight is given to the year when profits were more) and for a given p_t, w_t decreases as t increases (more weightage is given to recent observations than older ones).