A New Approach for Construction and Estimation of Prior

Jae-Young Kim

School of Economics
Seoul National University

July, 2017
KEA-APEA Conference 2017
Korea University

About Prior

- This paper is about 'prior' for statistical inference:
  - a new approach to construction of prior
  - new usage of prior
- Prior information is regarded as non-sample, or before-sample information.
- Prior could contain information on any unknowns or unobservables for statistical inference and forecasting:
  - prior for parameters given a parametric model.
  - prior for models in model selection and averaging
  - and more

Prior for Parametric Inference

- Given a density $f(x|\theta)$ and prior $\pi(\theta)$, the posterior is obtained:

$$\pi(\theta|x) = \frac{f(x|\theta)\pi(\theta)}{m_\pi(x)}$$

where

$$m_\pi(x) = \int f(x|\theta)\pi(\theta)d\theta$$

is the marginalized density.

- Prior and sample information are combined to yield posterior

$$\begin{pmatrix} \text{posterior} \\ \textit{after-data} \\ \textit{belief} \end{pmatrix} \propto \begin{pmatrix} \text{likelihood} \\ \textit{sample} \\ \textit{information} \end{pmatrix} \times \begin{pmatrix} \text{prior} \\ \textit{before-data} \\ \textit{belief} \end{pmatrix}$$

Literature on Determination of Prior

- Subjective priors: De Groot (1970), de Finetti (1972), Lindley (1982)
- Conjugate priors
- Non-informative priors: Laplace (1812), Jeffrey (1961), Hartigan (1964), Jaynes (1968, 1983), Villegas (1977, 1981, 1984)
- Empirical Bayes: Robbins (1956), Carlin and Louis (1996), Gholami et al. (2015), Efron (2015).

Prior Information

- Prior may contain all the relevant, direct and indirect, information on the unknowns.
- Prior is for *weighing* different possibilities of unknowns to make the best possible decision or inference.
- As such, prior should depend on model specification, direct versus indirect information, sampling anomalies (randomness), alternative rules for decision making, etc.

Prior - Extended Concept and Usage

- In the presence of heterogeneity: Heterogeneity may exist across cross-sectional or time-series domain. Use indirect evidence from other (concurrent) related samples. e.g. a medical example in a later section
- In the presence of specification uncertainty: Use the fact that [the likelihood and prior combined] matches the marginal density of observations.
- In the presence of gap between two disciplines: The frequentist confidence statement and the Bayesian probability statement may not be the same.
- Application for meta analysis: In meta analysis different statistics may contain different amount of information for the unknown

Deconvolution (1)

- $X_i, i = 1, \cdots, n$, is a set of observables from a parametric family of densities $f_\theta(\cdot) \equiv f(\cdot, \theta)$ for $\theta \in \Theta \subset I\!R^k$.
- An (unknown) prior density $\pi(\theta)$ has produced $\tilde{\theta}$ with a realized value $\theta \in \Theta$. Then, X_i is produced from $f(x|\theta)$.
- The marginal density of x_i from π, $m_\pi(x_i)$, is defined as

$$m_\pi(x_i) = \int f(x_i|\theta)\pi(\theta)d\theta.$$

- An inverse problem to obtain π from m_π can be established.
- $x_i,\ i = 1, ..., n$ could be the past data or the current data, the former case being called empirical Bayes and latter compound decision problem. (Robbins (1951,1956))

Deconvolution (2)

- From the definition of marginalized density, $m_\pi(x_i)$,

$$m_\pi(x_i) = \int f(x_i|\theta)\pi(\theta)d\theta,$$

  we consider an inverse problem of estimating π from m_π.
- Estimating π from m_π is the problem of deconvolution, the "Bayes deconvolution", which is a well known ill-posed problem. (Efron (2015))
- Also, m is often not directly available.
- We discuss a convenient approach to estimating $\pi(\theta)$.

Alternative Approach (1)

- Given the marginal density,

$$m_\pi(x) = \int f(x|\theta)\pi(\theta)d\theta,$$

take logarithms on both sides

$$\log(m_\pi(x)) = \log \int f(x|\theta)\pi(\theta)d\theta.$$

- Take expectations of $\log(m_\pi(x))$ with respect to the (true) density of x, $m(x)$:

$$E^m[\log(m_\pi(x))] = \int m(x) \log(m_\pi(x))dx.$$

Alternative Approach (2)

- Define a measure of discrepancy between m and m_π

$$d(m, m_\pi) = E^m \left[\log \frac{m(x)}{m_\pi(x)} \right]$$
$$= E^m[\log(m(x))] - E^m[\log(m_\pi(x))] \qquad (1)$$

  which is the Kullback-Leibler distance (relative entropy distance) between m and m_π.
- We know that

$$d(m, m_\pi) \geqslant 0, \ = 0 \ \text{ iff } \ m = m_\pi$$

- Our solution for π is the one that minimizes $d(m, m_\pi)$.
- Notice that the first term on the RHS of (1) has nothing to do with π, which implies the following useful result.

Lemma

Let $m_\pi(x)$ be as defined in (1). Then, it is true that

$$\underset{\pi}{\operatorname{argmin}}(d(m, m_\pi)) = \underset{\pi}{\operatorname{argmax}}(E^m[\log(m_\pi(x))]) \tag{2}$$

- From the above lemma, our solution for π is the one that maximizes $E^m[\log(m_\pi(x))]$ in (1):

$$\begin{aligned} E^m[\log(m_\pi(x))] &= \int m(x) \log(m_\pi(x)) dx \\ &= \int m(x) \log\left(\int f(x|\theta)\pi(\theta)d\theta\right) dx. \end{aligned} \tag{3}$$

Specification for π

- We consider $\pi(\theta)$ satisfying certain moment conditions. Consider the following problem:

$$\max_{\pi} \int \pi(\theta) \log \pi(\theta) d\theta$$

$$\text{subject to } \int \theta^r \pi(\theta) d\theta = \mu_r{}', \; r = 1, \ldots, p. \tag{4}$$

  That is, we consider $\pi(\theta)$ as the maximum entropy density subject to a set of moment conditions.

- The max-entropy density reflects the full uncertainty about π: Only a set of *sure* information, the moment conditions, are taken care of.
- Thus, it is the most conservative density available, given the sure information.

- It is well known that the solution $\pi^{me}(\theta)$ of this problem is

$$\pi^{me}(\theta) = \frac{\exp(\sum_r \lambda_r \theta^r)}{\int \exp(\sum_r \lambda_r \theta^r) d\theta} = \exp\left(\sum_{r=1}^{p} \lambda_r \theta^r + \lambda_0\right) \qquad (5)$$

where $\lambda_0 = -\log\left(\int \exp(\sum_{r=1}^{p} \lambda_r \theta^r) d\theta\right)$.

Alternative Approach (3)

- Now, with π^{me} in place of π, we have

$$E^m\left[\log(m_{\pi^{me}}(x))\right] = \int m(x)\log\left(\int f(x|\theta)\exp\left[\sum_{r=0}^{p}\lambda_r\theta^r\right]d\theta\right)dx. \quad (6$$

- Let λ^* be the solution for $\lambda = (\lambda_0, \cdots, \lambda_p)'$:

$$\lambda^* = \underset{\lambda}{\operatorname{argmax}} \int m(x)\log\left(\int f(x|\theta)\exp\left[\sum_{r=0}^{p}\lambda_r\theta^r\right]d\theta\right)dx. \quad (7)$$

Empirical Version

- Let $\Theta_J \subset \Theta$ for $\Theta_J = \{\theta_j : j = 1, \ldots, J\}$, a discrete subset of Θ.
- Given a sample x_i, $i = 1, \cdots, n$, and Θ_J, we define the empirical objective function:

$$E^{\hat{m}}[\log(\hat{m}_{\pi^{me}}(x))] = \sum_{i=1}^{n} \frac{1}{n} \log \left(\sum_{j=1}^{J} f(x_i|\theta_j) \exp \left[\sum_{r=0}^{p} \lambda_r {\theta_j}^r \right] \right). \quad (8)$$

- Define the solution $\tilde{\lambda}_{n,J}$ of the empirical optimization:

$$\tilde{\lambda} \quad = \quad \underset{\lambda}{\operatorname{argmax}} \, E^{\hat{m}}[\log(\hat{m}_{\pi^{me}}(x))]. \quad (9)$$

Asymptotics

- Let $\hat{P}^{\pi^{me}}_{n,J}$ be the probability measure induced by $\hat{m}_{\pi^{me}}$ with a sample of size n and Θ_J:

$$\hat{P}^{\pi^{me}}_{n,J}(t) = \sum_{i=1}^{n} \hat{m}_{\pi^{me}}(x_i)\mathbf{1}_{(x_i \leqslant t)}(t).$$

- It is natural to consider the relation between $\hat{P}^{\pi^{me}}_{n,J}(t)$ and $P^{\pi^{me}}(t)$, the 'true' probability measure defined by

$$P^{\pi^{me}}(t) = \int_{-\infty}^{t} m_{\pi^{me}}(x)dx$$

- Indeed, $\hat{P}^{\pi^{me}}_{n,J}(t)$ converges to $P(t)^{\pi^{me}}$.

Theorem 1

Assume that $\{X_i\}$ is a stationary and ergodic process. Assume that $f(\cdot|\theta)$ is continuous in x for every θ, that $f(x|\cdot)$ is continuous in θ for every x, and that $\pi(\cdot)$ is continuous in θ. Let λ^* be the unique solution of the problem (7) and $\hat{\lambda}_{n,J}$ be that of the problem (9). Then we have

$$\hat{\lambda}_{n,J} \xrightarrow{p} \lambda^* \quad \text{as} \quad n, J \to \infty.$$

Theorem 2

Under the same conditions as in Theorem 1, it is true that

$$\lim_{n,J\to\infty} Q\Big(\sup_{t\in\mathbb{R}} |\hat{P}^{\pi^{me}}_{n,J}(t) - P^{\pi^{me}}(t)| > \epsilon \Big) = 0$$

where Q is the probability measure of the sample implied by $m(x)$, $Q(t) = P^{\pi^{me}}(t) = \int_{-\infty}^{t} m(x)dx$.

Algorithms

- Grid search method is slow and inefficient but most reliable with a reasonable dimension of grids for a model with a low dimensional parameter space.
- Nelder-Mead direct search algorithm may be a reasonable option, which does not involve gradient information.
  - Computationally inefficient, but it works successfully for wide class of problems.
  - Starts with a simplex of $n + 1$ vertices in the search region of $\mathbb{R}^n$.
  - Then the algorithm transforms the simplex along the surface by reflection, expansion, contraction or shrinkage for each iteration step.

Application (1)

Example 1: Misspecification

$$f(\cdot|\theta): \quad X_i \sim N(\theta, 1)$$
$$\pi(\theta): \quad \theta \sim N(1, 0.5)$$

$m(x)$ is obtained from

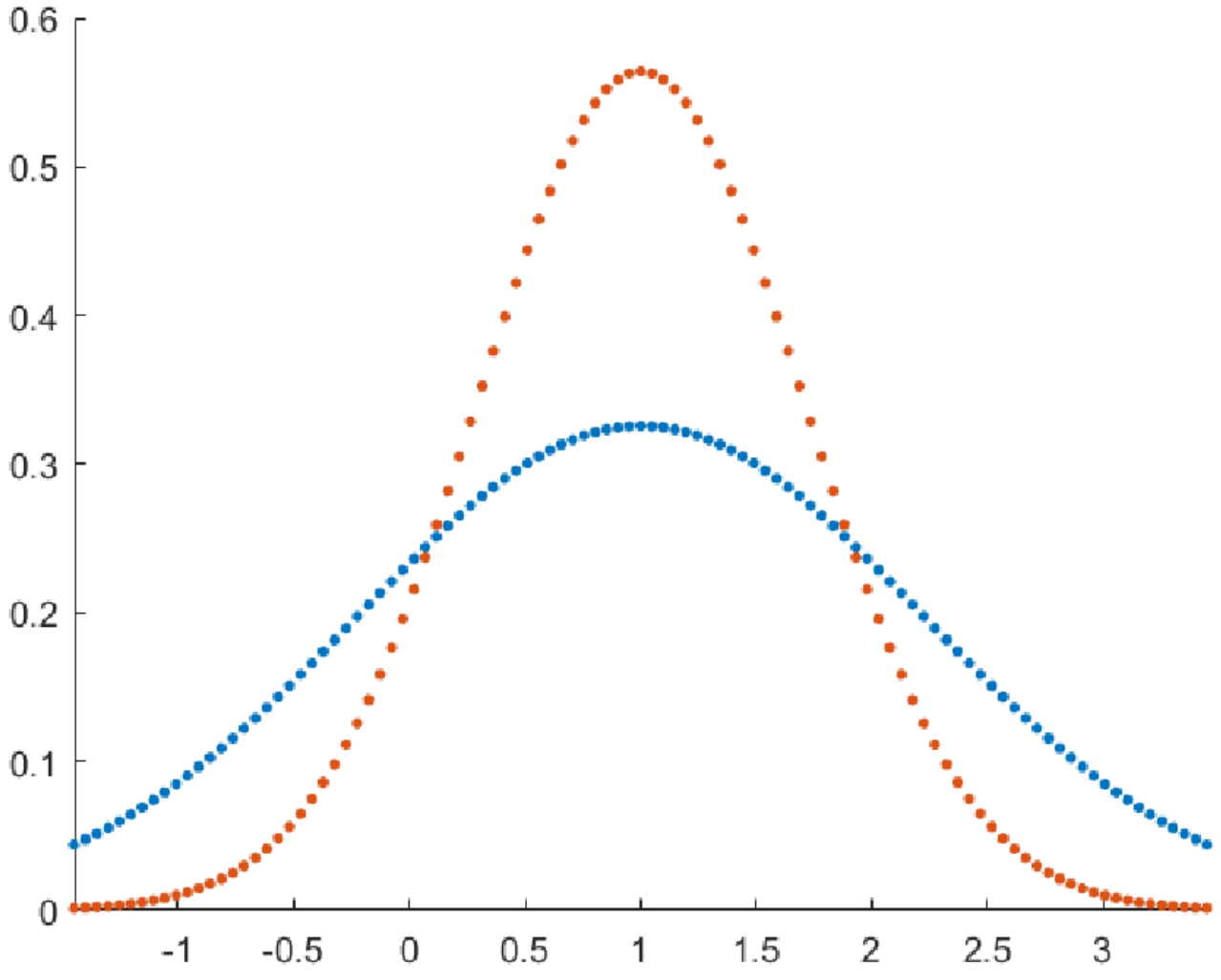
$$\hat{m}(x) = \sum_{j=1}^{J} f(x|\theta_j)\pi(\theta_j).$$

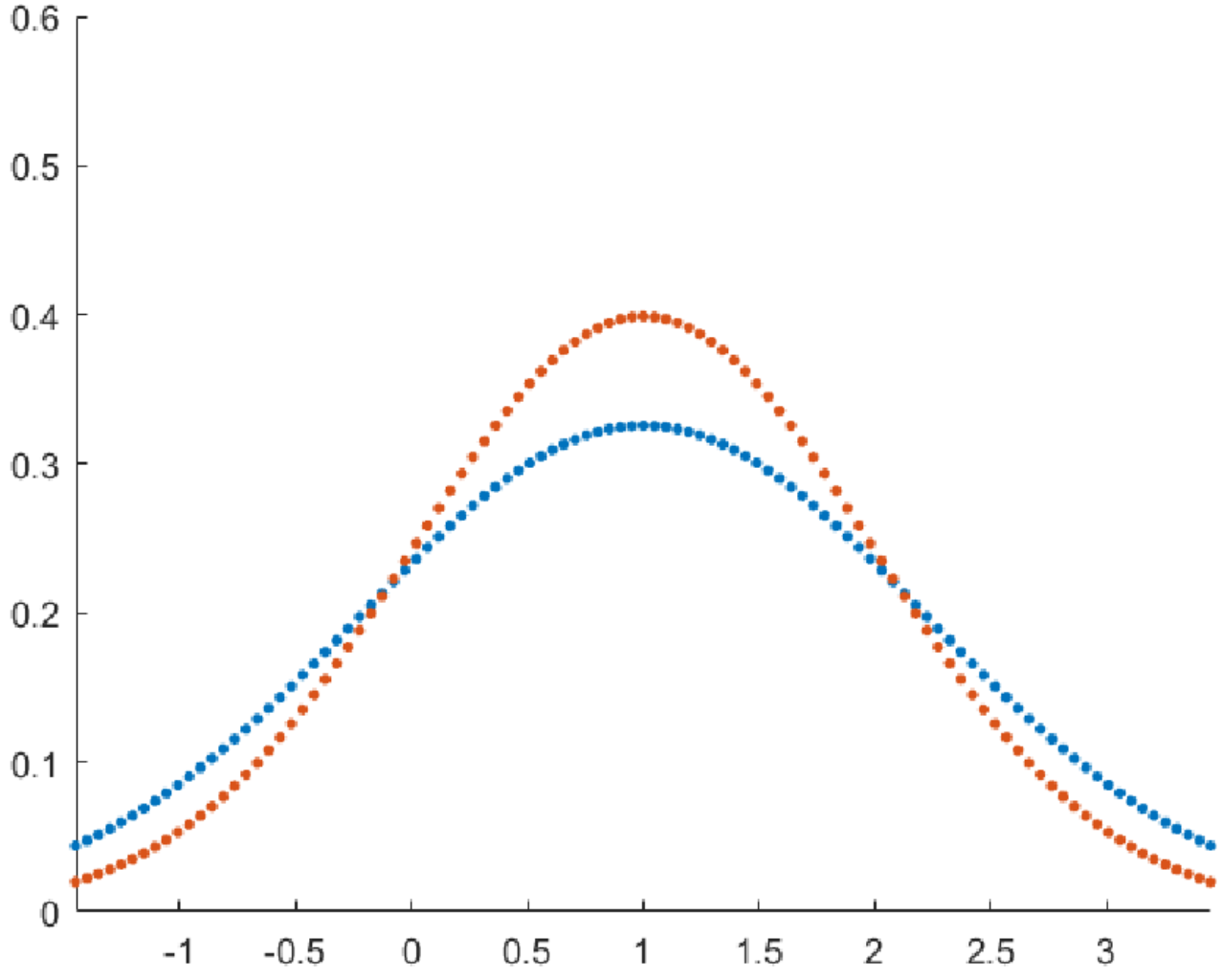
Get $x_i,\ i = 1, ..., n$ from $\hat{m}(x)$. Estimate λ and π from (9).
Figure 4: Get $\hat{\pi}$, given the true f

$$f(\cdot|\theta): \ X_i \sim N(\theta, 1).$$

Figure 5: Get $\hat{\pi}$, given a wrong f

$$f(\cdot|\theta): \ X_i \sim N(\theta, 0.5).$$


0.6
0.5
0.4
0.3
0.2
0.1
0
-1
-0.5
0
0.5
1
1.5
2
2.5
3


0.6
0.5
0.4
0.3
0.2
0.1
0
-1
-0.5
0
0.5
1
1.5
2
2.5
3

Application (2)

Example 2

Cancer surgery involving the removal of surrounding lymph nodes.

Data: (Data set from Seoul National University Hospital)
N= 548 surgeries (individuals), each reporting

$n = \#(\text{nodes removed})$ and $x = \#(\text{nodes found positive})$.

(Results to be included.)

Conclusions (1)

- We discuss a new approach to the estimation of prior based on minimization of relative entropy applied to the Bayes deconvolution.
- Estimation could done either by the past data or by the current data.
- Estimation by the current data brings us to a new area of compound decision problem, idea of which is due to Robbins (1951,1955).
- Can use indirect evidence from other (concurrent) related samples to get a prior and posterior which brings a more sensible inference basis, e.g. a medical example.

Conclusions (2)

- Provide a natural mitigator of the problem of model misspecification, using the fact that the likelihood and prior combined matches the marginal density of observations.
- A new inference method for detecting misspecification can be explored.
- May study a framework for finite sample analysis in the frequentist approach.