Imprecise Two-Stage Maximum Likelihood Estimation

Lev Utkin

Munich, September 2009

Initial statistical data

1. We have a set of observations $\mathbf{X} = (x_1, ..., x_n)$, for instance, the successive intervals between failures.
2. $x_1, ..., x_n$ are a realization of random variables $X_1, ..., X_n$. The r.v. X_i is governed by a pdf $p_i(x \mid \mathbf{b}_i, \mathbf{d})$ with vectors of parameters $\mathbf{b}_i, \mathbf{d}$.
3. It is assumed that there exists a function $f(i, \mathbf{b}, \mathbf{d})$ such that the vector $\mathbf{b}_i$ completely depends on the number i and the vectors of parameters $\mathbf{b}$, $\mathbf{d}$ through the function f, i.e., $\mathbf{b}_i = f(i, \mathbf{b}, \mathbf{d})$.

A standard way for computing the parameters

The likelihood function is

$$L(\mathbf{X} \mid \mathbf{b}, \mathbf{d}) = \Pr\{X_1 = x_1, ..., X_n = x_n\}$$
$$= \prod_{i=1}^{n} p_i(x_i \mid \mathbf{b}_i, \mathbf{d}).$$

Values of the parameters **b**, **d** should be chosen in such a way that makes $L(\mathbf{K} \mid \mathbf{b}, \mathbf{d})$ achieve its maximum.

Problems we could meet

1. A large number of parameters and the small amount of statistical data:
    - it is difficult to estimate the actual impact of every parameter;
    - it is difficult to compute the optimal values of parameters.
2. The precise distribution or pdf p_i might be unknown. We can say only about some set of distributions $\mathcal{M}_i$ due to:
    - the limited amount of statistical data.

The first obvious idea (1)

The first obvious idea following from the second problem: Every X_i is governed by an unknown CDF belonging to a set $\mathcal{M}_i(\mathbf{d})$ depending on a vector of parameters $\mathbf{d}$ and defined by **lower and upper CDFs**:

$$\underline{F}_i(x \mid \mathbf{d}) = \inf_{F(x)\in\mathcal{M}_i(\mathbf{d})} F(x),\ \overline{F}_i(x \mid \mathbf{d}) = \sup_{F(x)\in\mathcal{M}_i(\mathbf{d})} F(x).$$

The first obvious idea (2)

IMPORTANT:

1. $\mathcal{M}_i(\mathbf{d})$ is the set of *all* CDFs bounded by $\underline{F}_i(k \mid \mathbf{d})$ and $\overline{F}_i(k \mid \mathbf{d})$, so it is *not* the set of parametric distributions having the same parametric form as the bounding distributions.
2. $\mathcal{M}_i(\mathbf{d})$ depends on $\mathbf{d}$.
3. We can not now maximize of the standard likelihood function over parameters. What can we do?

The second idea: maximum of the likelihood function over the set of CDFs

1. The standard likelihood function is the joint probability which has to be maximized over sets of parameters. But we have a set of probabilities. Therefore, we choose the largest probability in the set, i.e., we maximize the likelihood function over the set of probabilities depending on $\mathbf{d}$.
2. **Let us fix the parameters $\mathbf{d}$.**
3. The likelihood function $L(\mathbf{X} \mid \mathbf{d}, F)$ is maximized over all distributions F from $\mathcal{M}_i(\mathbf{d})$ and the resulting "modified" likelihood function depends on $\mathbf{d}$:

$$L^*(\mathbf{X} \mid \mathbf{d}) = \max_{F \in \mathcal{M}_1(\mathbf{d}), \dots, F \in \mathcal{M}_n(\mathbf{d})} L(\mathbf{X} \mid \mathbf{d}, F).$$

The third idea: maximum of the “modified” likelihood function over the set of parameters d

By assuming that the “modified” likelihood function $L^*(\mathbf{X} \mid \mathbf{d})$ depends on $\mathbf{d}$, we maximize it over the set of $\mathbf{d}$ in order to find $\mathbf{d}$, i.e.,

$$L^*(\mathbf{X} \mid \mathbf{d}) \to \max_{\mathbf{d}}.$$

Returning to the second idea: maximum of the likelihood function over the set of CDFs

- In other words, we have to find optimal distribution functions in every $\mathcal{M}_i(\mathbf{d})$ which *can* depend on $\mathbf{d}$.
- How to find them?

The maximized likelihood function (discrete case)

Proposition

If random variables $X_1, ..., X_n$ are independent and discrete, then there holds

$$\max_{\mathcal{M}} \Pr\{X_1 = x_1, ..., X_n = x_n\} = \prod_{i=1}^{n} \{\overline{F}_i(x_i) - \underline{F}_i(x_i - 1)\}.$$

"Precise" case

Corollary

If $\overline{F}_i(x) = \underline{F}_i(x) = F_i(x)$*, then*

$$\max_{\mathcal{M}} \Pr\left\{X_1 = x_1, ..., X_n = x_n\right\} = \prod_{i=1}^{n} p_i(x_i) = L(\mathbf{X} \mid \mathbf{d}).$$

Here $p_i(k)$ *is the probability mass function corresponding to the distribution function* $F_i(k)$*.*

We have the standard likehood function.

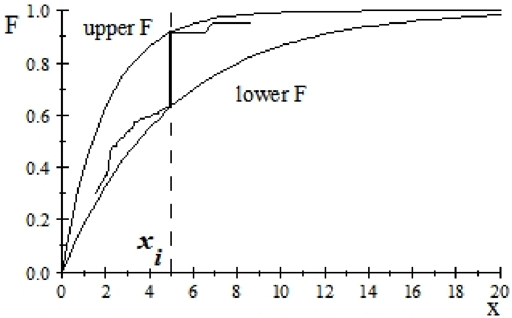
The maximized likelihood function (continuous case)

Proposition

If random variables $X_1, ..., X_n$ are independent and continuous, then there holds

$$\max_{\mathcal{M}} \Pr \{X_1 = x_1, ..., X_n = x_n\} = \prod_{i=1}^{n} \left\{ \overline{F}_i(x_i) - \underline{F}_i(x_i) \right\} \delta(x_i).$$

Optimal distribution function (continuous case)



The maximized likelihood function (the lack of independence)

Proposition

If there is no information about independence of random variables $X_1, ..., X_n$, then there holds

$$\max_{\mathcal{M}} \Pr\{X_1 = x_1, ..., X_n = x_n\} = \min_{i=1,...,n} \left\{ \overline{F}_i(x_i) - \underline{F}_i(x_i - 1) \right\}.$$

“Precise” case

Corollary

If $\overline{F}_i(x) = \underline{F}_i(x)$, *then*

$$\max_{\mathcal{M}} \Pr\left\{X_1 = x_1, ..., X_n = x_n\right\} = \min_{i=1,...,n} p_i(x_i).$$

We have the possibilistic likehood function.

Returning to the third idea: maximum of the "modified" likelihood function over the set of parameters d

$$L^{*}(\mathbf{X} \mid \mathbf{d}) = \bigotimes_{i=1}^{n} \left\{ \overline{F}_i(x_i \mid \mathbf{d}) - \underline{F}_i(x_i - 1 \mid \mathbf{d}) \right\} \to \max_{\mathbf{d}} .$$

Here the operator $\bigotimes$ can be $\prod$ (independence) or min (unknown interaction).

The next problem is how to construct the lower and upper CDFs

Three obvious methods can be proposed:

1. Using the imprecise Bayesian models.
2. Using the method of moments.
3. Using the confidence intervals on the mean and variance.

The imprecise Bayesian inference models

1. Imprecise Dirichlet model (Walley 1996);
2. Imprecise models for inference in exponential families (Quaeghebeur and de Cooman 2005).

The lower and upper CDFs for $\mathcal{M}_i(\mathbf{d})$ are constructed by means of **an imprecise Bayesian model** conditioned on the parameters $\mathbf{d}$ and the function $f(i, \mathbf{b}, \mathbf{d})$. The parameters $\mathbf{b}$ are replaced by **caution parameter** s or parameters s_1, s_2. The **imprecision** of the model is defined by the caution parameter s.

The imprecise method of moments (1)

By having k moments, we can restrict a set of probability distributions (or pdfs) by the constraints:

$$\mathbb{E}(x^i) = m_i(\mathbf{d}),\ \ i = 1, ..., k,$$

or

$$\sum\nolimits_{j=1}^{N} p(v_j) v_j^i = \frac{1}{n} \sum\nolimits_{j=1}^{n} x_j^i,\ \ i = 1, ..., k,$$

or

$$\int_{-\infty}^{\infty} v^i p(v) \mathrm{d}v = \frac{1}{n} \sum\nolimits_{j=1}^{n} x_j^i,\ \ i = 1, ..., k.$$

Here $p \in \mathcal{M}$. In other words, the set of sample moments produces the set $\mathcal{M}$.

The imprecision is defined by a number of moments.

The imprecise method of moments (2)

The parametric (with parameters **d**) linear programming:

$$\underline{F}(x \mid \mathbf{d}) = \min_{p} \sum_{j=1}^{N} p(v_j) I_{(-\infty,x]}(v_j),$$

$$\overline{F}(x \mid \mathbf{d}) = \max_{p} \sum_{j=1}^{N} p(v_j) I_{(-\infty,x]}(v_j),$$

subject to

$$\sum_{j=1}^{N} p(v_j) v_j^i = m_i(\mathbf{d}), \ \ i = 1, ..., k.$$

In regression models: $x_j = y_j - f(\mathbf{x}_j, \mathbf{d})$ and:

$$m_i(\mathbf{d}) = \frac{1}{n} \sum_{j=1}^{n} \left(y_j - f(\mathbf{x}_j, \mathbf{d})\right)^i.$$

The imprecise method of moments (3)

The parametric (with parameters **d**) linear programming:

$$\underline{F}(x \mid \mathbf{d}) = \min_{p} \int_{-\infty}^{\infty} I_{(-\infty,x]}(v)p(v)\mathrm{d}v,$$

$$\overline{F}(x \mid \mathbf{d}) = \max_{p} \int_{-\infty}^{\infty} I_{(-\infty,x]}(v)p(v)\mathrm{d}v,$$

subject to

$$\int_{-\infty}^{\infty} v^{i}p(v)\mathrm{d}v = m_i(\mathbf{d}), \ i = 1, ..., k.$$

In regression models: $x_j = y_j - f(\mathbf{x}_j, \mathbf{d})$ and:

$$m_i(\mathbf{d}) = \frac{1}{n}\sum_{j=1}^{n} \left(y_j - f(\mathbf{x}_j, \mathbf{d})\right)^{i}.$$

The imprecise method of moments (4): example - Chebyshev's inequality

We take only two moments and obtain Chebyshev's inequality. Bounds for the CDF are

$$\underline{F}(t \mid \mathbf{d}) = \begin{cases} 1 - \dfrac{m_2(\mathbf{d}) - m_1^2(\mathbf{d})}{(m_1(\mathbf{d}) - t)^2 + m_2(\mathbf{d}) - m_1^2(\mathbf{d})}, & t \geq m_1(\mathbf{d}) \\ 0, & t < m_1(\mathbf{d}) \end{cases},$$

$$\overline{F}(t \mid \mathbf{d}) = \begin{cases} \dfrac{m_2(\mathbf{d}) - m_1^2(\mathbf{d})}{(m_1(\mathbf{d}) - \tau)^2 + m_2(\mathbf{d}) - m_1^2(\mathbf{d})}, & t \leq m_1(\mathbf{d}) \\ 1, & t > m_1(\mathbf{d}) \end{cases}.$$

Confidence intervals on the mean and variance

95% confidence intervals on the mean and variance ($\alpha = 0.05$):

$$[\underline{m}_1(\mathbf{d}), \overline{m}_1(\mathbf{d})] = \left[m_1(\mathbf{d}) - \frac{t_{\alpha/2,N-1}\hat{\sigma}(\mathbf{d})}{\sqrt{N}},\ m_1(\mathbf{d}) + \frac{t_{\alpha/2,N-1}\hat{\sigma}(\mathbf{d})}{\sqrt{N}} \right],$$

$$[\underline{\sigma}^2(\mathbf{d}), \overline{\sigma}^2(\mathbf{d})] = \left[\frac{(N-1)\hat{\sigma}^2(\mathbf{d})}{\chi^2{}_{\alpha/2,N-1}},\ \frac{(N-1)\hat{\sigma}^2(\mathbf{d})}{\chi^2{}_{1-\alpha/2,N-1}} \right],$$

$$\underline{F}(x \mid \mathbf{d}) = \min\left\{ \Phi\left((x - \overline{m}_1(\mathbf{d}))/\overline{\sigma}(\mathbf{d}) \right), \Phi\left((x - \overline{m}_1(\mathbf{d}))/\underline{\sigma}(\mathbf{d}) \right) \right\},$$
$$\overline{F}(x \mid \mathbf{d}) = \max\left\{ \Phi\left((x - \underline{m}_1(\mathbf{d}))/\overline{\sigma}(\mathbf{d}) \right), \Phi\left((x - \underline{m}_1(\mathbf{d}))/\underline{\sigma}(\mathbf{d}) \right) \right\}.$$

The imprecision is defined by α. In regression models:

$$\hat{\sigma}^2(\mathbf{d}) = \frac{1}{n} \sum_{j=1}^{n} \left(y_j - f(\mathbf{x}_j, \mathbf{d}) \right)^2 - \left(\frac{1}{n} \sum_{j=1}^{n} \left(y_j - f(\mathbf{x}_j, \mathbf{d}) \right) \right)^2.$$

Returning to the third idea: maximum of the "modified" likelihood function over the set of parameters d

Now the "modified" likelihood function has been defined

$$L^{*}(\mathbf{X} \mid \mathbf{d}) = \bigotimes_{i=1}^{n} \left\{ \overline{F}_i(x_i \mid \mathbf{d}) - \underline{F}_i(x_i - 1 \mid \mathbf{d}) \right\} \rightarrow \max_{\mathbf{d}} .$$

Questions

?