# MAXIMIZING PROBIT LIKELIHOOD FUNCTIONS USING MAJORIZATION

JAN DE LEEUW AND JEFFREY LEWIS

ABSTRACT. This abstract will be an abstract.

## 1. PROBLEM

In many estimation problems, especially in the social and behavioural sciences, we have to minimize a deviance function of the form

$$\mathcal{D}(\tau,\theta)=-2\sum_{i=1}^n\sum_{s=1}^p y_{is}\log\left[\Phi(\tau_{s+1}+\eta_i(\theta))-\Phi(\tau_s+\eta_i(\theta))\right]. \tag{1}$$

Here $\Phi$ is the cumulative standard normal, $\tau$ is a sequence of thresholds, which satisfies

$$-\infty=\tau_1<\tau_2<\cdots<\tau_p<\tau_{p+1}=+\infty,$$

and the $y_{is}$ are the observed choices, coded as indicators. Thus the $y_{is}$ are binary and satisfy

$$\sum_{s=1}^p y_{is}=1$$

for all $i=1,\cdots,n$. The thresholds $\tau$ can be either known, partially known, or unknown, depending on the application. In our examples we shall discuss both the function (1), and some variations of it which can be optimized by the same techniques.

---

*Date*: January 14, 2012.

2000 *Mathematics Subject Classification.* 62-04,62P20,62P25.

*Key words and phrases.* Probit models, choice models, item reponse theory, majorization algorithms.

Because we are dealing with normal distributions, it is convenient to introduce some shorthand notation. Let

$$\phi_{\mu\sigma}(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\},$$

and for $\phi_{0,1}(x)$ we simply write $\phi(x)$. In the same way $\Phi_{\mu\sigma}(z)=\int_{-\infty}^{z}\phi_{\mu\sigma}(x)$ and $\Phi_{0,1}(z)$ is simply $\Phi(z)$. Thus

$$\begin{aligned}\phi_{\mu,\sigma}(x)&=\frac{1}{\sigma}\phi(\frac{x-\mu}{\sigma}),\\ \Phi_{\mu,\sigma}(z)&=\Phi(\sigma z+\mu).\end{aligned}$$

## 2. EXAMPLES

See **??** for extensive discussion of models leading to likelihood functions of this form.

*Example* 2.1 (Discrete Normal). Suppose $x_1,\cdots,x_n$ are realizations of a random variable $\underline{x}$ with a discrete normal distribution with parameters $\mu$ and $\sigma$. This means

$$\mathbf{prob}(\underline{x}=s)=\Phi_{\mu,\sigma}(\tau_{s+1})-\Phi_{\mu,\sigma}(\tau_s).$$

Observe that we follow the convention of underlining random varibles [**?**]. In many cases, because of the inherent discreteness of actual observations, this is a more realistic model than the usual continuous normal one. The negative log-likelihood is

$$\begin{aligned}\mathcal{D}(\mu,\sigma,\tau)&=-\sum_{i=1}^{n}\sum_{s=1}^{p}y_{is}\log\left[\Phi(\sigma\tau_{s+1}+\mu)-\Phi(\sigma\tau_s+\mu)\right]=\\&=-\sum_{s=1}^{p}n_s\log\left[\Phi(\sigma\tau_{s+1}+\mu)-\Phi(\sigma\tau_s+\mu)\right]\end{aligned}$$

Here the $n_s$ are the frequencies in the intervals. In this example the thresholds are assumed known, if they were unknown we could always attain perfect fit.

*Example* 2.2 (Discrete Normal Regression). Now suppose $x_i$ is a realization of $\underline{x}_i$, which is discrete normal with mean $\eta_i(\theta)$ and variance $\sigma^2$. The negative log-likelihood becomes

$$\mathcal{D}(\tau,\theta,\sigma)=-\sum_{i=1}^n\sum_{s=1}^p y_{is}\log\left[\Phi(\sigma\tau_{s+1}+\eta_i(\theta))-\Phi(\sigma\tau_s+\eta_i(\theta))\right].$$

In the most common case, of course, $\eta_i(\theta)=x_i'\theta$, and we have the discrete linear regression model. Observe that in this case we can fit models with both $\tau$ is known, which means we still have to estimate $\sigma$, or with $\tau$ is unknown, in which case we absorb $\sigma$ into $\tau$.

*Example* 2.3 (Truncated Normal). Suppose we have a random variable $\underline{x}$ which takes values in the open interval $(\tau_0,\tau_1)$ and has density

$$\mathbf{prob}(x)=\frac{\phi_{\mu\sigma}(x)}{\Phi_{\mu,\sigma}(\tau_1)-\Phi_{\mu,\sigma}(\tau_0)}.$$

The negative log-likelihood is

$$\mathcal{D}(\mu,\sigma)=-\sum_{i=1}^n\log\phi_{\mu,\sigma}(x_i)+n\log\left[\Phi(\sigma\tau_1+\mu)-\Phi(\sigma\tau_0+\mu)\right]$$

In this case $\tau_0$ and $\tau_1$ will usually be known. There is a regression version with

$$\begin{aligned}(2)\quad\mathcal{D}(\theta,\sigma)=n\log\sigma-\sum_{i=1}^n\log\phi(\frac{x_i-\eta_i(\theta)}{\sigma})+\\+\sum_{i=1}^n\log\left[\Phi(\sigma\tau_1+\eta_i(\theta))-\Phi(\sigma\tau_0+\eta_i(\theta))\right]\end{aligned}$$

*Example* 2.4 (Censored Normal). Suppose we have a random variable $\underline{x}$ which takes values in the closed interval $[\tau_0,\tau_1]$. On the open interval $(\tau_0,\tau_1)$ the density is $\phi_{\mu,\sigma}$. In addition

$$\begin{aligned}\mathbf{prob}(\underline{x}=\tau_0)&=\Phi_{\mu,\sigma}(\tau_0),\\\mathbf{prob}(\underline{x}=\tau_1)&=1-\Phi_{\mu,\sigma}\tau_1.\end{aligned}$$

Thus the remaining normal mass is concentrated at the endpoints.

The negative log-likelihood, in the regression form, is

$$\mathcal{D}(\theta,\sigma)=-\sum_{\tau_0<x_i<\tau_1}\log\phi_{\eta_i(\theta),\sigma}(x_i)-\sum_{x_i=\tau_0}\log\Phi(\sigma\tau_0+\eta_i(\theta))-\sum_{x_i=\tau_1}\log\left[1-\Phi(\sigma\tau_1+\eta_i(\theta))\right].$$

Tobit Regression [**?**] is the special case in which $\tau_0=0$ and $\tau_1=+\infty$.

*Example* 2.5 (Item Response Theory).

## 3. ALGORITHMS

## 4. MAJORIZATION

All our algorithms are based on the following key lemma. It generalizes **??**.

**Lemma 4.1.** *Suppose* $-\infty\leq\alpha<\beta\leq+\infty$ *and define*

$$f(x)=-\log\left[\Phi(\beta+x)-\Phi(\alpha+x)\right].$$

*Then* $0<f''(x)<1$.

*Proof.* By direct calculation

$$f'(x)=-\frac{\phi(\beta+x)-\phi(\alpha+x)}{\Phi(\beta+x)-\Phi(\alpha+x)}$$

and

$$f''(x)=-\frac{(\beta+x)\phi(\beta+x)-(\alpha+x)\phi(\alpha+x)}{\Phi(\beta+x)-\Phi(\alpha+x)}+\left[\frac{\phi(\beta+x)-\phi(\alpha+x)}{\Phi(\beta+x)-\Phi(\alpha+x)}\right]^2$$

Now consider a doubly truncated standard normal random variable $\underline{x}$, truncated on the right at $\beta+x$ and on the left at $\alpha+x$. Using a formula from **?**, p. 158, we see that the variance of $\underline{x}$ is given by

$$\mathbf{V}(\underline{x})=1-f''(x).$$

But this immediately implies the bounds given in the theorem. □

By taking $\alpha$ or $\beta$ to inifinity, we obtain the previous result that $f(x) = -\log \Phi(x)$ satisfies $0 < f''(x) < 1$.

Lemma 4.1 generalizes the well-known result that $f$ is strictly convex.

We now use Lemma 4.1 to construct a majorization function for the deviance (1). First define

$$g_{is}(\tau, \theta) = -\frac{\phi(\tau_{s+1} + \eta_i(\theta)) - \phi(\tau_s + \eta_i(\theta))}{\Phi(\tau_{s+1} + \eta_i(\theta)) - \Phi(\tau_s + \eta_i(\theta))}$$

and

$$h_i(\tau, \tilde{}) = \sum_{s=1}^{p} y_{is} g_{is}(\tau, \theta).$$

**Theorem 4.2.**

$$\mathcal{D}(\tau, \theta) \leq \mathcal{D}(\tau, \tilde{\theta}) + \sum_{i=1}^{n} h_i(\tau, \tilde{\theta})(\eta_i(\theta) - \eta_i(\tilde{\theta})) + \frac{1}{2}\sum_{i=1}^{n}(\eta_i(\theta) - \eta_i(\tilde{\theta}))^2$$

*Proof.* □

## 5. ALGORITHM

$$\gamma(\tau, \theta) = \eta_i(\theta) - h_i(\tau, \theta)$$

$$\mathcal{S}(\tau, \theta, \tilde{\theta}) = \sum_{i=1}^{n}(\eta_i(\theta) - \gamma(\tau, \tilde{\theta})))^2$$

Discussion: correction for continuity.

## 6. FITTING THRESHOLDS

## REFERENCES

D. Böhning. The Lower Bound Method in Probit Regression. *Computational Statistics and Data Analysis*, 30:13–17, 1999.

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA, LOS ANGELES, CA 90095-1554

DEPARTMENT OF POLTICAL SCIENCE, UNIVERSITY OF CALIFORNIA, LOS ANGELES, CA 90095-????

*E-mail address*, Jan de Leeuw: deleeuw@stat.ucla.edu

*URL*, Jan de Leeuw: http://gifi.stat.ucla.edu

*E-mail address*, Jeffrey Lewis: jblewis@ucla.edu

*URL*, Jeffrey Lewis: http://www.polisci.ucla.edu/menu/people/faculty/jeffrey_lewis.htm