Sur la matrice d'information de Fisher dans le calcul du maximum de vraisemblance, avec des applications en modles de choix discrets

Fabian Bastin
Special thanks to Emma Frejinger (University of Montreal),
Mai Tien (University of Montreal),
Michel Toulouse (Vietnamese-German University)

MIAT, INRA, Toulouse, 10 avril 2015

Université de Montréal

CIRRELT

Problem context

Many estimation models can be expressed as a general stochastic program

$$\min_{\beta} g(\beta) \overset{def}{=} E[f(y,\beta)] \qquad \text{(true problem)}$$

The computation of this expectation requires an infinite population. In practice, we only have access to a finite number of observations, leading to the approximation

$$\min_{\beta} \hat{g}_N(\beta) \overset{def}{=} \frac{1}{N}\sum_{n=1}^{N}[f(y_n,\beta)], \qquad \text{(SAA problem)}$$

where y_n is the observational vector associated to observation n.

The previous program can be seen as a special application of sample average approximation (SAA) technique.

Examples

Least-squares:

$$\min_{\beta} \frac{1}{N} \sum_{n=1}^{N} \|f(x_n, \beta) - y_n\|^2.$$

Maximum likelihood:

$$\max_{\beta} \frac{1}{N} \sum_{n=1}^{N} \ln f(y_n, \beta).$$

Despite the logarithm operator, problems are very similar. We here focus on maximum likelihood, but many arguments can be applied to least-squares problems.

Note: the factor $\frac{1}{N}$ is often ignored.

Maximum likelihood estimation (MLE): solve

$$\max_{\beta} \widehat{LL}_N(\beta) = \frac{1}{N}\sum_{n=1}^{N} \ln f(y_n|\beta) \tag{1}$$

- $f(Y|\beta)$: some probability density function (pdf), defined on Y, conditioned on a set of parameters β,
- $y_1, \ldots, y_N$ are given observations.

(1) is the sample average approximation of the “true” problem

$$\max_{\beta} LL(\beta) = \mathbb{E}_Y[\ln f(y|\beta)]. \tag{2}$$

f does not necessarily correspond to the density of Y over the population, in which case the model is said to be misspecified.

Convergence of solutions

Under some regularity conditions, when N rises to infinity,

$$d(\hat{S}_N^*, S^*) \to 0 \text{ almost surely,}$$

where d is a distance measure, $\hat{S}_N^*$ and S^* are the sets of first-order critical points of (1) and (2), respectively, assuming that $\hat{S}_N^*$ and S^* are not empty (see e.g. Shapiro [9] and Shapiro, Dentcheva, Ruszczyński [10], Chapter 5).

Moreover, if these sets are singletons, we denote by $\hat{\beta}_N^*$ the solution of (1) and by β^* the solution of (2). We then have that

$$\sqrt{N}(\hat{\beta}_N^* - \beta^*) \Rightarrow \mathcal{N}(0, \Psi),$$

where $\Rightarrow$ designs the convergence in distribution, and $\mathcal{N}$ refers to the normal distribution.

Convergence of solutions (singletons)

Setting the gradient of the (true) log-likelihood to zero, it can be shown that

$$\Psi = H(\beta^*)^{-1} I(\beta^*) H(\beta^*)^{-1},$$

where

- $H(\beta^*) = \mathbb{E}_Y[\nabla^2_{\beta\beta} f(Y, \beta^*)]$,
- $I(\beta^*) = \mathbb{E}_Y\left[\frac{\nabla_\beta f(Y,\beta^*) \nabla_\beta f(Y,\beta^*)^T}{f^2(Y,\beta^*)}\right]$.

- $\nabla_\beta \ln f(Y, \beta)$: score
- $I(\beta)$: Fisher information matrix

(see e.g. Newey and McFadden [8])

Variance-covariance

The asymptotic variance-covariance can therefore be estimated by

$$\mathrm{Cov}(\hat{\beta}^*_N) = \frac{[H_N(\hat{\beta}^*_N)]^{-1} I_N(\hat{\beta}^*_N)[H_N(\hat{\beta}^*_N)]^{-1}}{N}$$

where

$$H_N(\hat{\beta}^*_N) = \frac{1}{N}\sum_{n=1}^{N} \nabla^2_{\beta\beta} \ln f(y_n, \hat{\beta}^*_N)$$

and

$$I_N(\hat{\beta}^*_N) = \frac{1}{N}\sum_{n=1}^{N} \frac{\nabla_\beta \ln f(y_n, \hat{\beta}^*_N) \nabla_\beta \ln f(y_n, \hat{\beta}^*_N)^{T}}{f(y_n, \hat{\beta}^*_N)}$$

are the sample average estimates of the Hessian and the information matrix, respectively.

Information matrix equality

Theorem

If the model is well specified, i.e. $f(Y,\beta^*)$ is the density of Y over the population, the Fisher information matrix is equal to the opposite of the Hessian matrix,

$$I(\beta^*) = -H(\beta^*).$$

The robust variance-covariance of the maximum likelihood estimator then becomes

$$-\frac{[H_N(\hat{\beta}^*_N)]^{-1}}{N}.$$

Therefore, the variances and covariances of the estimators are directly related to the curvature of the log-likelihood.

Pseudo-likelihood maximization

The information matrix equality still holds if f is not the true density but $f(y, \beta^*) = g(y)$ where $g(\cdot)$ is the density of Y.

However, this condition is not easy to guarantee, and even if the pseudo-maximum likelihood estimators are consistent, the information matrix equality can be violated.

Example (White [11]) We consider the estimation of the mean and variance of i.i.d. random variables Y_i. We assume $Y_i \sim N(\mu_0, \sigma_0^2)$. The quasi-log-likelihood of an observation is

$$
\begin{aligned}
\log f(Y_t, \mu, \sigma^2) &= log \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(Y_t-\mu)^2}{2\sigma^2}} \\
&= -log\sqrt{2\pi} - \log\sigma - \frac{(Y_t-\mu)^2}{2\sigma^2}.
\end{aligned}
$$

Pseudo-likelihood maximization (2)

If Y has nonzero variance and finite fourth moment, $\mu^* = \mu_0$ and $\sigma^* = \sigma_0$.

$$E[H(\mu_0, \sigma_0^2)] = \begin{pmatrix} -\frac{1}{\sigma_0^2} & 0 \\ 0 & -\frac{1}{2\sigma_0^4} \end{pmatrix} \qquad I[\mu_0, \sigma_0^2] = \begin{pmatrix} \frac{1}{\sigma_0^2} & \frac{\sqrt{\gamma_1}}{2\sigma_0^3} \\ \frac{\sqrt{\gamma_1}}{2\sigma_0^3} & \frac{\gamma_2 - 1}{4\sigma_0^4} \end{pmatrix}$$

$$E[H(\mu_0, \sigma_0^2)]^{-1} I[\mu_0, \sigma_0^2] E[H(\mu_0, \sigma_0^2)]^{-1} = \begin{pmatrix} \sigma_0^2 & \sqrt{\gamma_1}\sigma_0^3 \\ \sqrt{\gamma_1}\sigma_0^3 & (\gamma_2 - 1)\sigma_0^4 \end{pmatrix}$$

$$\text{(skewness)} \qquad \sqrt{\gamma_1} = \frac{E[(Y_t - \mu_0)^3]}{\sigma_0^3}$$

$$\text{(kurtosis)} \qquad \gamma_2 = \frac{E[(Y_t - \mu_0)^4]}{\sigma_0^4}.$$

Optimization problem

Is it possible to take advantage of the information matrix equality during the estimation process?

We face the optimization problem:

$$\max_{\boldsymbol{\beta}} LL(\boldsymbol{\beta}).$$

Assumption in this talk: the problem is unconstrained.

More generally, we consider the problem

$$\min_{x \in \mathcal{R}^n} f(x),$$

where we assume that $f : \mathcal{R}^n \to \mathcal{R} \in C^2$.

Newton method: introduction

If f is C^2, we can write the Taylor expansion of order 2:

$$f(x+d) \approx f(x) + \nabla_x f(x)^T d + \frac{1}{2} d^T \nabla^2_{xx} f(x) d.$$

Note: assume $\frac{1}{2} d^T \nabla^2_{xx} f(x^*) d > 0$ if $\|d\| < \epsilon$ (sufficient second-order criticality condition - local strong convexity).

If $\nabla_x f(x^*) = 0$, then x^* is a local minimizer.

Otherwise, we can choose d such that $\nabla_x f(x^*)^T d < 0$ and x^* is not a local minimizer: descent direction.

Newton method

At iteration k, define (around the current iterate x_k)

$$m_k(d) = f(x_k) + \nabla_x f(x_k)^T d + \frac{1}{2} d^T \nabla^2_{xx} f(x_k) d.$$

If $\nabla^2_{xx} f(x_k)$ is positive definite, $m_k(d)$ is convex in a neighborhood of x_k and we can minimize $m_k(d)$ by computing d such that $\nabla_d m_k(d) = 0$. In other terms, we compute d_k as

$$d_k = -[\nabla^2_{xx} f(x_k)]^{-1} \nabla_x f(x_k),$$

and we set

$$x_{k+1} = x_k + d_k.$$

If the starting point x^0 is close to the solution, the convergence is quadratic, but if the starting point is not good enough, the method can diverge!

Quasi-Newton method

It is often difficult to obtain an analytical expression for the derivatives, and they may be costly to compute at each iteration, so we prefer to turn to approximations.

First-order derivatives can be computed by finite differences:

$$\frac{\partial f}{\partial x_{[i]}} \approx \frac{f(x+\epsilon e_i) - f(x)}{\epsilon},$$

for small ϵ, and $e_i = \begin{pmatrix} 0 & 0 & \dots & 0 & 1 & 0 & \dots 0 \end{pmatrix}^T$ is the i^{th} canonical vector, or central differences:

$$\frac{\partial f}{\partial x_{[i]}} \approx \frac{f(x+\epsilon e_i) - f(x-\epsilon e_i)}{2\epsilon}.$$

This however requires $O(n)$ evaluations of the objective.

Hessian approximation: BFGS

We can approximate the Hessian with a similar approach, but the computation cost is then of $O(n^2)$, wich is usually too expensive.

An alternative is to construct an approximation of the Hessian that we will improve at each iteration, using the gained information. We then speak of quasi-Newton method.

Popular approaches are based on the secant condition

$$H_{k+1} d_k = w_k$$

where $d_k = x_{k+1} - x_k$ and $w_k = \nabla_x f(x_{k+1}) - \nabla_x f(x_k)$.

Other classical updates

A popular approximation is the BFGS (Broyden, Fletcher, Goldfarb and Shanno):

$$B_{k+1} = B_k + \frac{w_k w_k^T}{w_k^t d_k} + \frac{B_k d_k (B_k d_k)^T}{d_k^T B_k d_k}.$$

Another possible choice in trust-region is the symmetric rank-one (SR1)

$$H_{k+1} = H_k + \frac{(w_k - H_k d_k)(w_k - H_k d_k)^T}{(w_k - H_k d_k)^T d_k}$$

BFGS is always positive definite, not SR1!

Globalization of the Newton method

Global convergence: the algorithm must converge for any starting point. BUT it still converges to a local mimimum, not a global minimum!

So how to ensure global convergence? Globalization of the Newton method:

- linesearch methods;
- trust-region methods.

Line-search methods generate the iterates by setting

$$x_{k_1} = x_k + \alpha_k d_k.$$

where d_k is a search direction and $\alpha_K d_k$ is chosen so that

$$f(x_{k+1}) < f(x_k).$$

Linesearch methods are therefore descent methods, and a special case is the steepest descent method.

Trust region methods

Principle: at iteration k, approximately minimize a model m_k of the objective inside a region $\mathcal{B}_k$. A typical choice for m_k is

$$m_k(x_k + s) = f(x_k) + s^T \nabla f(x_k) + \frac{1}{2} s^T H_k s.$$

H_k: approximation of $\nabla^2_{xx} f(x_k)$.

We therefore have to solve the subproblem

$$\min_s m_k(x_k + s), \text{ such that } x_k + s \in \mathcal{B}_k.$$

The solution is the candidate iterate with candidate step s_k.

Computing the following ratio:

$$\rho_k = \frac{f(x_k + s_k) - f(x_k)}{m_k(x_k + s_k) - m_k(x_k)}.$$

Trust region methods (2)

Let η_1 and η_2 be constants such that $0 < \eta_1 \leq \eta_2 < 1$ (for instance, $\eta_1 = 0.01$ and $\eta_2 = 0.75$).

- If $\rho_k \geq \eta_1$, accept the candidate.
- If $\rho_k \geq \eta_2$, enlarge $\mathcal{B}_k$, otherwise reduce it or keep it the same.
- If $\rho_k < \eta_1$, reject the candidate and reduce $\mathcal{B}_k$.

Stop when some criteria are met (e.g. norm of the relative gradient must be small enough).

The neighborhood where we considere the model as valid is mathematically defined by a ball centered at x_k, and with a radius Δ_k:

$$\mathcal{B}_k = \{x \mid \|x - x_k\|_k \leq \Delta_k\}.$$

BHHH

Various alternatives to BFGS exist, as the Symmetric-Rank 1 (SR1) that is popular for nonconvex optimzation with trust-region.

For the maximum likelihood, we capitalize on the information equality:

$$H_N(\beta) \approx -I_N(\beta)$$

The BHHH method, proposed by Berndt, Hall, Hall, Hausman in 1974, simply replaces the Hessian by the opposite of the information matrix in the quasi-Newton method (or the information matrix in the minimization form).

Similar to Gauss-Newton for least-squares problems.

Gauss-Newton is known to work when residuals are small, but can experience issues otherwise.

Remedies

In the same way, the BHHH can lead to poor numerical performance if the model is not correctly specified.

Dennis and Schnabel [3] proposed to correct the Gauss-Newton approximation using standard secant Hessian approximations.

Bunch [2] applied the same idea to maximum likelihood estimation. While his approach supersedes the standard BHHH technique, it is often ignored.

We here focus on trust region (TR), while experiments have also been done with line search (LS) methods.

Combination of approximations

Remember we consider the problem

$$\max_{\beta} \frac{1}{N} \sum_{n=1}^{N} \ln f(y_n, \beta).$$

Then,

$$\begin{aligned}\nabla^2_{\beta\beta} \frac{1}{N} \sum_{n=1}^{N} \ln f(y_n, \beta) =& \frac{1}{N} \sum_{n=1}^{N} \frac{\nabla^2_{\beta\beta} f(y_n, \beta)}{f(y_n, \beta)} \\ &- \frac{1}{N} \sum_{n=1}^{N} \frac{\nabla_{\beta} f(y_n, \beta) \nabla_{\beta} f(y_n, \beta)^T}{f(y_n, \beta)^2},\end{aligned}$$

or

$$\nabla^2_{\beta\beta} \frac{1}{N} \sum_{n=1}^{N} \ln f(y_n, \beta) = -I_N(\beta) + \frac{1}{N} \sum_{n=1}^{N} \frac{\nabla^2_{\beta\beta} f(y_n, \beta)}{f(y_n, \beta)}.$$

Bunch's approximation

We approximate the Hessian as

$$\nabla^2_{\beta\beta} \frac{1}{N} \sum_{n=1}^{N} \ln f(y_n, \beta) \approx -I_N(\beta) + A_k.$$

It is possible to approximate the second term using some standard secant update.

- $A_k = 0 \rightarrow$: BHHH method.
- $A_k \neq 0 \rightarrow$: approximations combination.

We can switch between the two cases, depending on the performance of the model.

Bunch's switching method

Bunch considers one switch during during the optimization process, and does precise exactly when to proceed, except that one should start with BHHH and when close to the solution, switch to the corrected

Assuming that at iteration k the matrix H_k is available to approximate the next Hessian H_{k+1}, the new approximation can be obtained by specifying an appropriate secant condition

$$H_{k+1} d_k = w_k,$$

We can write

$$H_{k+1} = H_{BHHH}(\beta_{k+1}) + A_{k+1},$$

where A_{k+1} is an approximation of $A(\beta_{k+1})$.

Secant approximations

The secant equation can be rewritten as

$$(H_{BHHH}(\beta_{k+1}) + A_{k+1})d_k = w_k,$$

and by setting $\bar{w}_k^1 = w_k - H_{BHHH}(\beta_{k+1})d_k$, we obtain

$$A_{k+1}d_k = \bar{w}_k^1$$

A second secant equation is derived by approximating each individual Hessian matrix $\nabla^2_{\beta\beta} f(y_n, \beta)$:

$$\nabla^2_{\beta\beta} f(y_n, \beta_k)d_k \approx \nabla_\beta f(y_n, \beta_{k+1}) - \nabla_\beta f(y_n, \beta_k).$$

This gives

$$A(\beta_k)d_k \approx \frac{1}{N}\sum_{n=1}^{N} \frac{\nabla_\beta f(y_n, \beta_{k+1}) - \nabla_\beta f(y_n, \beta_k)}{f(y_n, \beta_k)}$$

So if we define $\bar{w}_k^2 = \frac{1}{N}\sum_{n=1}^{N} \frac{\nabla_\beta f(y_n,\beta_{k+1}) - \nabla_\beta f(y_n,\beta_k)}{f(y_n,\beta_k)}$, the second secant approximation is

$$A_{k+1}d_k = \bar{w}_k^2.$$

Switching ideas

We have many ways to approximate Hessian matrix.

- Secant method: BFGS.
- Secant method: SR1.
- Statistical approximation BHHH.
- Combining approximation (update A_k by BFGS or SR1).

Let $\mathcal{H}_k = \{H_k^{(i)}, i = 1, 2...\}$ be a set of Hessian approximations.

1. **Trust-region:** Consider different quadratic models

$$m_k^{(i)}(d) = g(x_k) + \nabla g_k^T d + \frac{1}{2} d^T H_k^{(i)} d, \quad H_k^{(i)} \in \mathcal{H}_k$$

   Choose one as the current sub-problem.

Model prediction (TR_{PRED})

At iteration k, we consider the set of Hessian approximation $\mathcal{H}_k$.

- The selected matrix (from previous iteration) is used for sub-problem.
- Suppose d_k is an approximate solution of the current subproblem. It the step is accepted, we predict the quadratic model for the next iteration by solving:

$$i^* = \arg\min_i ||m_k^{(i)}(d_k) - g(\beta_k + d_k)||.$$

- Update $H_{k+1}^{(i)} \in \mathcal{H}_{k+1}$.

Multiple subproblems (TR_{MULTI})

At iteration k:

- Compute $d_k^{(i)}$ is approximately solution of sub-problems:

$$\min_d \{g(\beta_k) + \nabla g_k^T d + \frac{1}{2} d^T H_k^{(i)} d, \quad H_k^{(i)} \in \mathcal{H}_k\}, \qquad H_k^{(i)} \in \mathcal{H}_k$$

- Suppose

$$i^* = \arg\max_i \left(g(\beta_k) - g(\beta_k + d_k^{(i)}) \right)$$

$\rightarrow$Step $d_k^{(i^*)}$ is chosen for current iteration.

This approach selects the subproblem providing the largest objective function reduction, but can be expensive, especially if $\mathcal{H}_k$ is large.

Application in discrete choice

An agent k has to make a choice among a discrete set of alternatives $\mathcal{A}_k$.

- Given a set of choice observations, how could be define a behavioral model?
- How to validate the model, and to make predictions?

We will focus here mainly on model formulation and estimation.

Main assumption: individuals act rationally, and aim to maximize their (perceived) utility.

Utility

Given $j \in \mathcal{A}_k$, the utility is to supposed to have the form

$$U_{j,k}(\boldsymbol{\beta}) = V_{j,k}(\boldsymbol{\beta}) + \epsilon_{j,k}.$$

where

- $V_{j,k}(\boldsymbol{\beta})$: deterministic part, that can be observed.
- $\epsilon_{j,k}(\boldsymbol{\beta})$: random part.
- $\boldsymbol{\beta}$: utility parameters.

This reflects that an external observer can only capture part of the utility, even if the utility is supposed to be perfectly known by the choice maker.

Consequence: only choice probabilities can be computed.

Random utility maximization

Therefore, the probability that individual k chooses j

$$P_{j,k}(\boldsymbol{\beta}) = P[U_{j,k}(\boldsymbol{\beta}) \geq U_{j,l}(\boldsymbol{\beta}), \forall\, l \in \mathcal{A}(k)]$$

or

$$P_{j,k}(\boldsymbol{\beta}) = P[V_{j,k}(\boldsymbol{\beta}) - V_{j,l}(\boldsymbol{\beta}) \geq \epsilon_{j,l} - \epsilon_{j,k}, \forall\, l \in \mathcal{A}(k)]$$

The explicit form of the random term $\epsilon_{j,k}$.

It is common to assume that $E[\epsilon_{j,k}] = 0$, while since only the difference of utilities matters, any constant expectation $E[\epsilon_{j,k}] = \alpha$ provides the same choice probabilities.

Logit model

One of the most famous models is the multinomial logit model (MNL).

Main advantage: analytical form.

The $\epsilon_{j,k}$ are supposed i.i.d. over j and k, and follow a Gumbel law (also known as Extreme Value Type I). Denoting the mean by μ and the scale factor by λ, the distribution function is

$$F(x) = e^{-e^{-\lambda(x-\mu)}}.$$

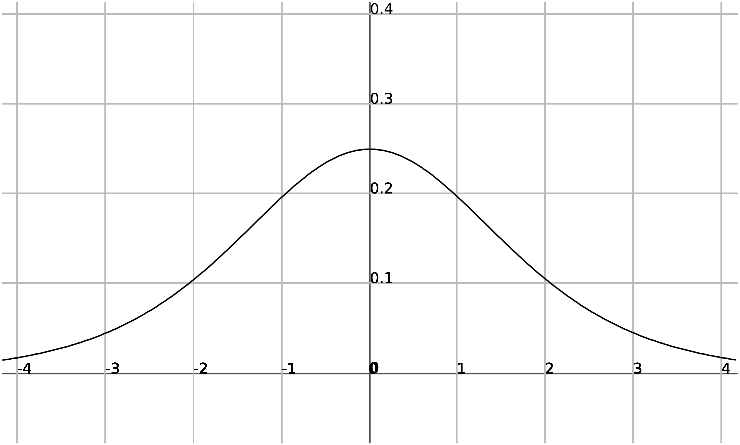
Logit model (cont'd)

Following the seminal work of McFadden (1973), we have

$$P_{j,k}(\beta) = \frac{e^{\lambda V_{j,k}}}{\sum_l e^{\lambda V_{j,l}}}.$$

Another justification is that if X and Y are Gumbel i.i.d., $Z = X - Y$ follows a logistic distribution of zero mean, et scale factor λ, which has a similar shape to a normal (but with heavier tails).

Logistic distribution



Substitution patterns

A fundamental property of logit models it the independence from irrelevant alternatives (I.I.A.). Consider the alternatives k_1 and k_2. Then,

$$\frac{P_{j,k_1}}{P_{j,k_2}} = \frac{e^{\lambda V_{j,k_1}}}{e^{\lambda V_{j,k_2}}} = e^{\lambda(V_{j,k_1} - V_{j,k_2})}.$$

In other words, the ratio of the choice probabilities of two alternatives does not depend of the other alternatives.

Not realistic if the set of alternatives contains alternatives with some degree of similarity.

For simplicity, we normalize the utilities to have $\lambda = 1$.

Red/blue buses example (Ben Akiva and Lerman, 1983)

Suppose that two transportation modes are available: car (c) or red bus (b_r), and that the choice probabilities are the same:

$$P_c = P_{b_r} = \frac{1}{2}.$$

Add the transportation mode blue bus (b_b). Then, from the I.I.A.,

$$P_c = P_{b_r} = P_{b_b} = \frac{1}{3}.$$

Note: the IIA issue can often be reduced with market segmentation, for instance if we could identify a population segment that favor red buses, while the other part prefer blue buses. But this can be tricky!

Alternatives sampling

If $\mathcal{A}_k$ is large, or even infinite, the computation of the choice probabilities can be difficult.

Due to the IIA, the model can nevertheless be consistently estimated on a sample of alternatives [6].

The probability that an individual j chooses alternative k given a sampled choice set $\mathcal{D}_j \subset \mathcal{A}_k$ is given by

$$P_{j,k}(\beta|D_n) = \frac{e^{V_{j,k}(\boldsymbol{\beta})+\ln \pi(\mathcal{D}_j|k)}}{\sum_{l\in\mathcal{D}_j} e^{V_{j,k}(\boldsymbol{\beta})+\ln \pi(\mathcal{D}_j|l)}} \tag{3}$$

$\ln \pi(\mathcal{D}_j|k)$ is the correction for sampling bias where $\pi(\mathcal{D}_j|k)$ is the probability of sampling choice set $\mathcal{D}_j$ given that k is the chosen alternative.

Alternatives sampling (cont'd)

The correction term is often ignored.

If $\pi(\mathcal{D}_k|j) = \pi(\mathcal{D}_k|l)$ for all $l \in \mathcal{D}_k$, this correction term can be safely discarded (as only the difference of utilities matters).

Otherwise, neglecting this term can lead to serious errors (see Frejinger [5], in the context of route choice).

Mixed Logit

The IIA issue leads to the developments on many other models that could relax this property, while maintaining an analytical solution: (cross-)nested logit, multivariate extreme value,. . .

The Mixed Multinomial Logit (MMNL), formaly introduced by McFadden and Train [7], became quite famous in transportation, and emerges now as popular in revenue management. The main reason for this success if the flexibility of the model, as (McFadden and Train, 2000)

> *Under mild regularity conditions, any discrete choice model derived from random utility maximization has choice probabilities that can be approximated as closely as one pleases by a MMNL model.*

Random-taste coefficients

The main idea is to let the parameters vector $\boldsymbol{\beta}$ to vary among individuals.

The parameters $\boldsymbol{\beta}$ is now a random vector over the population. An individual is associated to a specification realization $\boldsymbol{\beta}_j$, unknown by the observer.

The principle of random utility maximization still applies.

Conditionally to $\boldsymbol{\beta}_j$, j selects alternative k with probability

$$L_j(k, \boldsymbol{\beta}_j) = \frac{e^{V_{j,k}(\boldsymbol{\beta}_j)}}{\sum_{l \in \mathcal{A}(j)} e^{V_{j,l}(\boldsymbol{\beta}_j)}},$$

independently of other individuals.

Random-taste coefficients (cont'd)

Assume that

$$\boldsymbol{\beta}_q = h(\boldsymbol{\theta}, \mathbf{U})$$

where $\boldsymbol{\theta}$ is a vector of parameters and $\mathbf{U}$ is a given multivariate random vector (w.l.o.g. we can consider an uniform random vector on $(0,1)^s$.

The unconditional choice probability is

$$P_{j,k}(j, \boldsymbol{\theta}) = E[L_j(k, \boldsymbol{\beta})] = \int_{\mathcal{R}^s} L_j(k, \boldsymbol{\beta}) f_{\boldsymbol{\theta}}(\boldsymbol{\beta})\, d\boldsymbol{\beta}.$$

or

$$P_{j,k}(j, \boldsymbol{\theta}) = E[L_j(k, h(\boldsymbol{\theta}, \mathbf{U}))] = \int_{(0,1)^s} L_j(k, h(\boldsymbol{\theta}, \mathbf{u}))\, \d{\mathbf{u}}.$$

No more closed form for the choice probabilities!

Panel

The mixed logit can also explicitely account for sequence of observations from the same individual.

Assume that individual j delivers T_j observations, associate with a sequence of correlated realizations $\boldsymbol{\beta}_{j,t}$. The probability of this sequence is

$$L_j^{T_j}(k_1, \ldots, k_{T_j}, \boldsymbol{\beta}_{j,t_1}, \ldots, \boldsymbol{\beta}_{j,T_j}) = \prod_{t=1}^{T_j} L_j(k_t, \boldsymbol{\beta}_{j,t}),$$

It is common to represent the individual choices correlations by letting

$$\boldsymbol{\beta}_{j,t_1} = \boldsymbol{\beta}_{j,t_2} = \ldots = \boldsymbol{\beta}_{j,T_j} = \beta_j.$$

In other words, we represent taste variations among the population only.

Panel (cont'd)

The unconditional choice probability is now

$$P_{j,\mathbf{k}}(j, \boldsymbol{\theta}) = E[L_j^{T_j}(k, h(\boldsymbol{\theta}, \mathbf{U}))] = \int_{(0,1)^s} L_j^{T_j}(\mathbf{k}, h(\boldsymbol{\theta}, \mathbf{u}))\, \underset{\cdot}{\mathbf{u}}.$$

where $\mathbf{k} = (k_1, \ldots, k_{T_j})$.

Panel data have become popular in transportation research, especially in stated preference data, as they cost less.

Maximum likelihood

We consider estimation of the model parameters by maximizing the log-likelihood:

$$\max_{\theta} LL(\theta) = \frac{1}{m} \sum_{j=1}^{m} \ln P_{j,\mathbf{k}}(j, \boldsymbol{\theta})).$$

where m is the number of individuals.
The estimation can be done by any continuous optimization algorithm, but the problem is possibly nonconvex.

Issue for mixed logit: computation of choice probabilities.

Simulated maximum likelihood

The choice probabilities are approximated using sample average over **U**:

$$\hat{P}^{n}_{j,k}(\boldsymbol{\theta}) \approx \frac{1}{n} \sum_{a_j=1}^{n} L_j^{T_j}(\mathbf{k}, h(\boldsymbol{\theta}, u_{a_j})).$$

Statistical properties are easy to study in the Monte Carlo setting.

For finite n, the log-likelihood presents some simulation bias due to the logarithm operator.

The estimated parameters are asymptotically unbiased and consistent if n grows fast enough with m [1].

Mixed logit datasets

Tests reported for two real datasets for mixed logit, and one for route choice.

Mixed logit datasets:

1. Cybercal model (`SP2`): stated preferences dataset collected at the Baltimore/Washington (Cirillo and Xu, 2010). 8 factors: 3 constant, 1 lognormally distributed, 5 normally distributed.
2. Iris model (`IRIS`): stated preferences dataset collected in Brussels, 2002. 18 factors: 11 constant, 5 normally distributed, 2 normally or lognormally distributed.

Route choice dataset

Borlänge network which was used in Fosgerau et al.[4].

- 3077 nodes, 7459 links, 21452 link pairs.
- Travel times are assumed static and deterministic.
- 1832 trips corresponding to simple paths with a minimum of 5 links.
- 466 destinations, 1420 different origin-destination (OD) pairs and more than 37,000 link choices .

Tested using sampling of paths [5], without and with a path-size argument that aims to capture the correlation between paths.

Summary of datasets

Data set	SP2	IRIS	PS	PSL
observations	2466 (2740)	2602 (871)	1832 (1832)	1832 (1832)
alternatives	2	8	50	50
variables	9	19	4	5

Compared algorithms

[1] **TR-BHHH:** Trust region algorithm with the BHHH

[2] **TR-BFGS:** Trust region algorithm with the BFGS

[3] **TR-SR1:** Trust region algorithm with the SR1

[4] **TR-BUNCH**^{1}**:** Bunch's switching approach with BHHH^{corr1}-BFGS

[5] **TR-BUNCH**^{2}**:** Bunch's switching approach with BHHH^{corr2}-BFGS

[6] **TR-PRED:** Trust region algorithm with the predictive model

[7] **TR-MULTI:** Trust region algorithm with the multi-subproblems model

[8] **LS-BHHH:** Line search algorithm with the BHHH

[9] **LS-BFGS:** Line search algorithm with the BFGS

[10] **LS-PRED:** Line search algorithm with the predictive model

Comparison of algorithms: mixed logit

Algorithms		SP2	IN	ILN
Trust region	TR-BHHH	27.0 (27.0)	23.9 (23.9)	37.0^* (37.0) [9]
	TR-BFGS	52.9 (52.9)	155.1 (155.1)	147.6 (147.6)
	TR-SR1	42.1 (42.1)	241.5 (241.5)	238.4^*(238.4) [2]
	TR-BUNCH^1	20.6 (20.6)	33.9 (33.9)	57.4 (57.4)
	TR-BUNCH^2	20.9 (20.9)	34.5 (34.5)	57.6 (57.6)
	TR-PRED	**14.2** (14.2)	21.8 (21.8)	**54.7** (54.7)
	TR-MULTI	46.4 (23.2)	40.4 (20.2)	77.4 (38.4
Line search	LS-BHHH	28.1 (14.6)	**20.1** (17.6)	78.8 (46.2)
	LS-BFGS	31.8(15.8)	126.0(98.9)	202.5 (142.0)
	LS-PRED	34.7 (15.1)	20.5 (18.1)	70.5 (43.8)

In brackets: number of iterations.

In squared brackets: number of instance with failures.

Comparison of algorithms: route choice

Algorithms		PS	PSL
Trust region	TR-BHHH	40.5 (40.5)	58.2 (58.2)
	TR-BFGS	19.6 (19.6)	22.5 (22.5)
	TR-SR1	24.5 (24.5)	25.4 (25.4)
	TR-BUNCH[1]	51.3 (51.3)	51.0 (51.0)
	TR-BUNCH[2]	51.3 (51.3)	51.0 (51.0)
	TR-PRED	20.6 (20.6)	19.6 (19.6)
	TR-MULTI	33.2 (16.6)	31.4 (15.7)
Line search	LS-BHHH	22.6 (22.1)	22.2 (21.7)
	LS-BFGS	**19.0** (17.3)	**19.1** (17.6)
	LS-PRED	22.6 (22.1)	22.2 (21.7)

Number of switches

Algorithms		SP2	IN	ILN	PS	PSL
Trust region	TR-PRED	5.3	4.7	18.4	6.4	5.7
	TR-MULTI	6.5	7.8	11.5	8.4	7.7
Line search	LS-PRED	1.4	1.0	0.9	1.0	1.0

The trust-region framework is more adapted to the switching strategy.

Information matrix test

Reformulate the information matrix equality as

$$H(\beta^*) + I(\beta^*) = 0.$$

The sum can be consistently estimated by

$$I_N(\hat{\beta}_N^*) + H_N(\hat{\beta}_N^*).$$

White [11] designed a test statistic based on the jointly normally asymptotically distributed property of $D_N(\hat{\beta}_N^*) = I_N(\hat{\beta}_N^*) + H_N(\hat{\beta}_N^*)$

$$\sqrt{N} D_N^{\eta}(\hat{\beta}_N^*) \Rightarrow \mathcal{N}(0, \mathcal{V}_N(\hat{\beta}_N^*)).$$

Here we note that for a matrix A, vector A^{η} is defined by taking η indicators of interest in A.

Information matrix test (2)

An asymptotic χ^2 statistic test is

$$\wp_N = N D_N^\eta(\hat{\beta}_N^*)^T \mathcal{V}_N(\hat{\beta}_N^*)^{-1} D_N^\eta(\hat{\beta}_N^*) \Rightarrow \chi_\eta^2$$

where χ_η^2 is chi-square distribution with η degrees of freedom. The value of $D_N^\eta(\hat{\beta}_N^*)$ and $\mathcal{V}_N(\hat{\beta}_N^*)$ are defined by

$$\begin{aligned} D_N^\eta(\hat{\beta}_N^*) &= \frac{1}{N}\sum_{n=1}^{N} d_n^\eta(y_n|\hat{\beta}_N^*) \\ \mathcal{V}_N(\hat{\beta}_N^*) &= \frac{1}{N}\sum_{n=1}^{N} \left[\psi_n(\hat{\beta}_N^*)\psi_n(\hat{\beta}_N^*)^T\right] \end{aligned} \quad (4)$$

where

$$d_n(y_n|\hat{\beta}_N^*) = [\nabla_\beta \ln f(y_n|\hat{\beta}_N^*)][\nabla_\beta \ln f(y_n|\hat{\beta}_N^*)]^T + \nabla_{\beta\beta}^2(\ln f(y_n|\hat{\beta}_N^*))$$

and $\psi_n(\hat{\beta}_N^*) = d_n^\eta(y_n|\hat{\beta}_N^*) - \nabla_\beta D_N^\eta(\hat{\beta}_N^*) H_N(\hat{\beta}_N^*)^{-1} \nabla_\beta \ln f(y_n|\hat{\beta}_N^*)$.

Application: route choice

- Given a network defined in terms of links and nodes. A path is a sequence of links that connects an origin to a destination.
- Given a transportation mode and an origin-destination pair (O-D), what is the chosen path for going from O to D?

Probabilistic route choice analysis

- Given a sample of observations of origins, destinations and connecting paths.
- The objective is to formulate an econometric model for the choice of path conditional on origins and destinations.
- Assigning probabilities to paths over choice set a paths in a way that is consistent with rational behavior and that model parameters can be consistently estimated.
- Estimation of the parameters based on the observations.
- Discrete choice models typically used for this problem.

Challenges

1. Choice set is unknown (large size of choice set)
    - Classical approaches: Generating choice sets of paths.
    - Frejinger et al. (2009): Importance sampling of alternatives using sampling protocol (PL: Path Logit).
    - Fosgerau at al. (2013): link-based recursive logit (RL) with unrestricted choice set.
2. Path utilities may be correlated due to physical overlap in the network
    - Correction utilities: Path Size (PS), Link Size (LS),. . .
    - Nested, Mixed logit models,. . .

Simulated data

Number of observations	IM test			IIA test		
	RL	RL-LS	PL	RL	RL-LS	PL
18320	-	0	-	-	1	-
1832	0	1	0	0	2	0
500	0	3	0	2	2	3
100	3	7	3	6	9	5

Table: Number of type I errors (over 20 samples), 0.05 significance level

Number of observations	IM test		IIA test	
	RL	PL	RL	PL
1832	0	0	0	0
500	0	0	0	0
100	20	20	4	4

Table: Number of Type II errors (over 20 samples), 0.05 significance level

Real data

Model	IM test			IIA test		
	$\bar{\chi}^2_\eta$	η	p-value	$\bar{\chi}^2_\eta$	η	p-value
RL-LS	159.3	15	3.39e-26	74.8	5	1.02e-14
RL	89.16	10	7.86e-15	48.4	4	7.79e-10
PSL	148.3	15	5.24e-24	74.3	5	1.30e-14
PL	63.89	10	6.60e-10	40.1	4	4.13e-8

Table: Test statistic values for IM and IIA tests for models estimated on real data

Real data (cont'd)

Data	Model	Nb. of variables	Estimation		Tests	
			BHHH	BFGS	IM	IIA
Simulated	RL-LS	4	0.96	1.32	12.05	3.14
	RL	3	0.47	0.81	9.56	2.63
	PL	3	0.01	0.02	2.65	3.69
Real	RL-LS	5	2.91	2.42	17.74	3.99
	RL	4	1.98	1.60	9.68	2.72
	PSL	4	0.05	0.02	2.91	4.05
	PL	4	0.03	0.02	2.73	3.77

Table: Computational time, in hours, for estimation and testing

References I

Fabian Bastin, Cinzia Cirillo, and Philippe L. Toint.
Application of an adaptive Monte Carlo algorithm to mixed logit estimation.
Transportation Research Part B, 40(7):577–593, 2006.

David S. Bunch.
Maximum likelihood estimation of probabilistic choice models.
SIAM J. Sci. Stat. Comp., 8(1):56–70, 1987.

John E. Dennis Jr and Robert B. Schnabel.
Least change secant updates for quasi-Newton methods.
SIAM Review, 21(4):443–459, 1979.

Mogens Fosgerau, Emma Frejinger, and Anders Karlstrom.
A link based network route choice model with unrestricted choice set.
Transportation Research Part B, 56:70–80, 2013.

References II

Emma Frejinger, Michel Bierlaire, and Moshe Ben-Akiva.
Sampling of alternatives for route choice modeling.
Transportation Research Part B, 43(10):984–994, 2009.

Daniel L. McFadden.
Modelling the choice of residential location.
In A. Karlquist et al., editor, *Spatial Interaction Theory and Residential Location*, pages 75–96. North Holland, Amsterdam, The Netherlands, 1978.

Daniel L. McFadden and Kenneth Train.
Mixed MNL models for discrete response.
Journal of Applied Econometrics, 15(5):447–270, 2000.

References III

Whitney K. Newey and Daniel McFadden.
Large sample estimation and hypothesis testing.
In R.F. Engle and D.L. McFadden, editors, *Handbook of Econometrics*, volume IV, chapter 36, pages 2111–2245. Elsevier, Amsterdam, The Netherlands, 1986.

Alexander Shapiro.
Monte carlo sampling methods.
In Alexander Shapiro and Andrzej Ruszczyński, editors, *Stochastic Programming*, volume 10 of *Handbooks in Operations Research and Management Science*, chapter 6, pages 353–425. Elsevier, 2003.

References IV

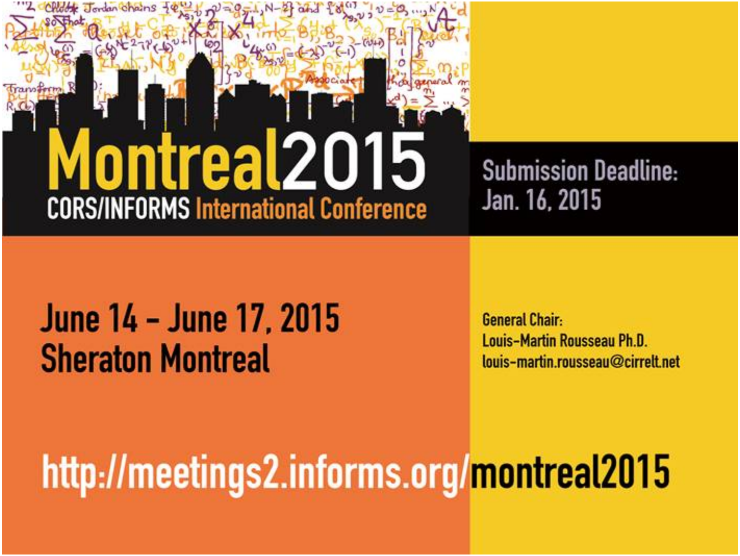
Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński.
Lectures on Stochastic Programming.
SIAM, Philadelphia, PA, USA, 2009.

Halbert White.
Maximum likelihood of misspecified models.
Econometrica, 50(1):1–25, 1982.

See you in Montreal?

Summer School on Dynamic Discrete Choice Models: Econometric Models and Operations Research Methods, June 10-12, 2015, Universit de Montral: https://symposia.cirrelt.ca/Summer-School/en

See you in Montreal!



Montreal2015

CORS/INFORMS International Conference

Submission Deadline:
Jan. 16, 2015

June 14 – June 17, 2015
Sheraton Montreal

General Chair:
Louis-Martin Rousseau Ph.D.
louis-martin.rousseau@cirrelt.net

http://meetings2.informs.org/montreal2015