EM algorithm requires repeated numerical optimization at each iteration

EM algorithm requires repeated numerical optimization at each iteration

Problem: No closed-form solution for weighted standard logit

Result: Extremely slow iterations, especially with many fixed coefficients

EM algorithm requires repeated numerical optimization at each iteration

Problem: No closed-form solution for weighted standard logit

Result: Extremely slow iterations, especially with many fixed coefficients

MM Algorithm Solution:

Uses quadratic lower bound to create surrogate function with closed-form solution

Newton's Method for optimization:

$$\beta^{m+1} = \beta^m - H^{-1}(\beta^m) \cdot \nabla L(\beta^m)$$

Newton's Method for optimization:

$$\beta^{m+1} = \beta^{m} - H^{-1}(\beta^{m}) \cdot \nabla L(\beta^{m})$$

where

- ∇ = Gradient vector (first derivatives)
- H = Hessian matrix (second derivatives)

JOURNAL OF APPLIED ECONOMETRICS
J. Appl. Econ. **32**: 841–857 (2017)
Published online 20 June 2016 in Wiley Online Library
(wileyonlinelibrary.com) DOI: 10.1002/jae.2532

MM ALGORITHM FOR GENERAL MIXED MULTINOMIAL LOGIT MODELS

JONATHAN JAMES*
Department of Economics, Orfalea College of Business, California Polytechnic State University, San Luis Obispo, CA, USA

SUMMARY

This paper develops a new technique for estimating mixed logit models with a simple minorization–maximization (MM) algorithm. The algorithm requires minimal coding and is easy to implement for a variety of mixed logit models. Most importantly, the algorithm has a very low cost per iteration relative to current methods, producing substantial computational savings. In addition, the method is asymptotically consistent, efficient and globally convergent. Copyright © 2016 John Wiley & Sons, Ltd.

Received 26 August 2014; Revised 8 March 2016

1. INTRODUCTION

Accommodating unobserved heterogeneity is critical for studying discrete response models and, more importantly, in using these models to construct counterfactuals. The mixed logit model is a flexible discrete-choice framework that allows the unobserved error structure of the utility maximization problem to be arbitrarily correlated across choices and time. In theory, the mixed logit model is extremely powerful, as McFadden and Train (2000) show that, with a flexible enough error structure, it can approximate any random utility model with an arbitrary degree of accuracy.

However, estimating mixed logit models is challenging because this more complicated error structure produces a difficult-to-maximize integrated likelihood function. Maximizing the likelihood with quasi-Newton methods is unattractive because they are either difficult to implement, being applicable to models only where it is realistic to derive and code problem-specific analytical gradients, or they have a high cost per iteration, either through the computation of the analytical gradient or conducting many evaluations of the likelihood to numerically approximate the gradient. An alternative approach for solving difficult maximization problems is to transfer optimization to simpler surrogate functions (Lange *et al.*, 2000). The use of surrogate functions is the engine for the expectation-maximization (EM) algorithm (Dempster *et al.*, 1977), a popular alternative to quasi-Newton methods which has been adopted for estimating mixed logit models (Bhat, 1997; Train, 2007, 2008). The EM algorithm forms a surrogate function that is an expected complete data likelihood. This makes the EM algorithm very easy to implement because maximizing the surrogate function only requires complete data maximum likelihood which often has a closed-form solution. Train (2007) illustrates this simplicity with an EM algorithm for estimating a mixed logit model with normally distributed random coefficients that entails iteratively computing a sample mean and a sample covariance.

While the EM algorithm is unmatched in its simplicity, its application to mixed logit models has a major shortcoming. In many situations, maximizing the surrogate function requires finding the maximum to a standard complete data logit model, which does not have a closed-form solution, requiring numerical optimization, in some cases multiple numerical optimizations, at each iteration.[1] This

* Correspondence to: Jonathan James, Department of Economics, Orfalea College of Business, California Polytechnic State University, San Luis Obispo, CA, USA. E-mail: jjames04@calpoly.edu

[1] In the latent class random coefficients model, each iteration of the EM algorithm requires solving a separate standard logit for each of the latent classes (Bhat, 1997; Train, 2007, 2008).

Problem: Multinomial logit Hessian depends on P's that change every iteration

$$H = \sum_i \sum_j P_{ij} x_{ij} x_{ij}'$$

Problem: Multinomial logit Hessian depends on P's that change every iteration

$$H = \sum_i \sum_j P_{ij} x_{ij} x_{ij}'$$

MM Insight: Replace probability weights with constants to get something like OLS

Problem: Multinomial logit Hessian depends on P's that change every iteration

$$H = \sum_i \sum_j P_{ij} x_{ij} x_{ij}'$$

MM Insight: Replace probability weights with constants to get something like OLS

Bound Matrix: $B = -\frac{1}{2} \sum_i \left[\sum_j x_{ij} x_{ij}' - \frac{1}{J} \left(\sum_j x_{ij} \right) \left(\sum_j x_{ij} \right)' \right]$

No probabilities! Just data (x_{ij}) like $X'X$ in OLS

Problem: Multinomial logit Hessian depends on P's that change every iteration

$$H = \sum_i \sum_j P_{ij} x_{ij} x_{ij}'$$

MM Insight: Replace probability weights with constants to get something like OLS

Bound Matrix: $B = -\frac{1}{2} \sum_i \left[\sum_j x_{ij} x_{ij}' - \frac{1}{J} \left(\sum_j x_{ij} \right) \left(\sum_j x_{ij} \right)' \right]$

No probabilities! Just data (x_{ij}) like $X'X$ in OLS

Result: Compute B^{-1} once instead of inverting P-dependent matrices every iteration

This is why MM is fast

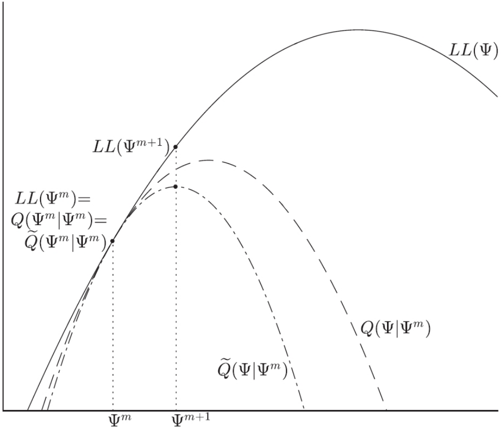

Figure 1. Sketch of MM algorithm

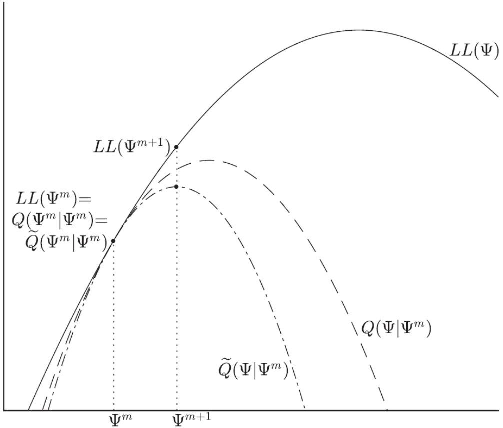

Figure 1. Sketch of MM algorithm

$LL(\Psi)$**:** Original log-likelihood

- Complicated, hard to maximize

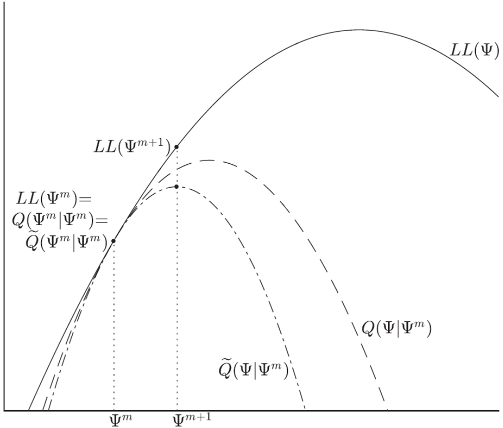

Figure 1. Sketch of MM algorithm

$LL(\Psi)$: Original log-likelihood

- Complicated, hard to maximize

$Q(\Psi|\Psi^m)$: EM surrogate

- Bounds LL from below
- Still requires numerical optimization

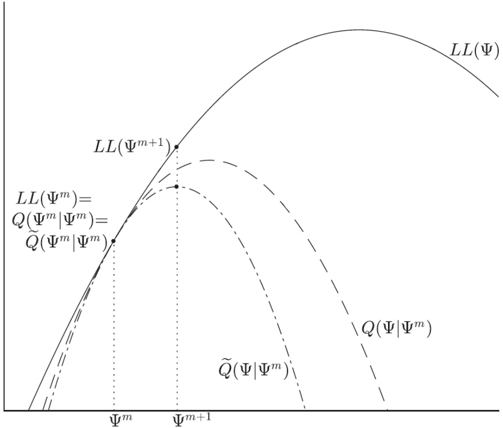

Figure 1. Sketch of MM algorithm

$LL(\Psi)$: Original log-likelihood

- Complicated, hard to maximize

$Q(\Psi|\Psi^m)$: EM surrogate

- Bounds LL from below
- Still requires numerical optimization

$\tilde{Q}(\Psi|\Psi^m)$: MM surrogate

- Bounds EM surrogate from below
- Has closed-form solution

All functions equal at Ψ^m, guaranteeing ascent to Ψ^{m+1}

MM Algorithm Advantages:

Computational Speed: 5-8x faster than EM in experiments

MM Algorithm Advantages:

Computational Speed: 5-8x faster than EM in experiments

Memory Efficiency: Uses sufficient statistics, no need to store simulated datasets

MM Algorithm Advantages:

Computational Speed: 5-8x faster than EM in experiments

Memory Efficiency: Uses sufficient statistics, no need to store simulated datasets

Parallelizable: Unlike EM, easily distributes computations

MM Algorithm Advantages:

Computational Speed: 5-8x faster than EM in experiments

Memory Efficiency: Uses sufficient statistics, no need to store simulated datasets

Parallelizable: Unlike EM, easily distributes computations

Simple Implementation: Much easier to code than analytical gradients

MM Algorithm Advantages:

Computational Speed: 5-8x faster than EM in experiments

Memory Efficiency: Uses sufficient statistics, no need to store simulated datasets

Parallelizable: Unlike EM, easily distributes computations

Simple Implementation: Much easier to code than analytical gradients

Competitive with quasi-Newton methods, especially for panel data

Most beneficial for:

- Models with many fixed coefficients
- Panel data settings
- Complicated models where analytical gradients are impractical

Most beneficial for:

- Models with many fixed coefficients
- Panel data settings
- Complicated models where analytical gradients are impractical

Limitations:

- May require more iterations than EM (especially large choice sets)
- Still requires numerical simulation
- Performance advantage diminishes in cross-sectional data