A Likelihood Approach to Fitting Nonlinear Mixed-Effects Models to Pharmacokinetic and Pharmacodynamic Data

Douglas Bates

University of Wisconsin - Madison

<Bates@Wisc.edu>

Slides for this presentation are available at
`lme4.R-forge.R-project.org/slides/`

Midwest Biopharmaceutical Statistics Workshop
Muncie, Indiana
May 26, 2010

Outline

Outline

Outline

Outline

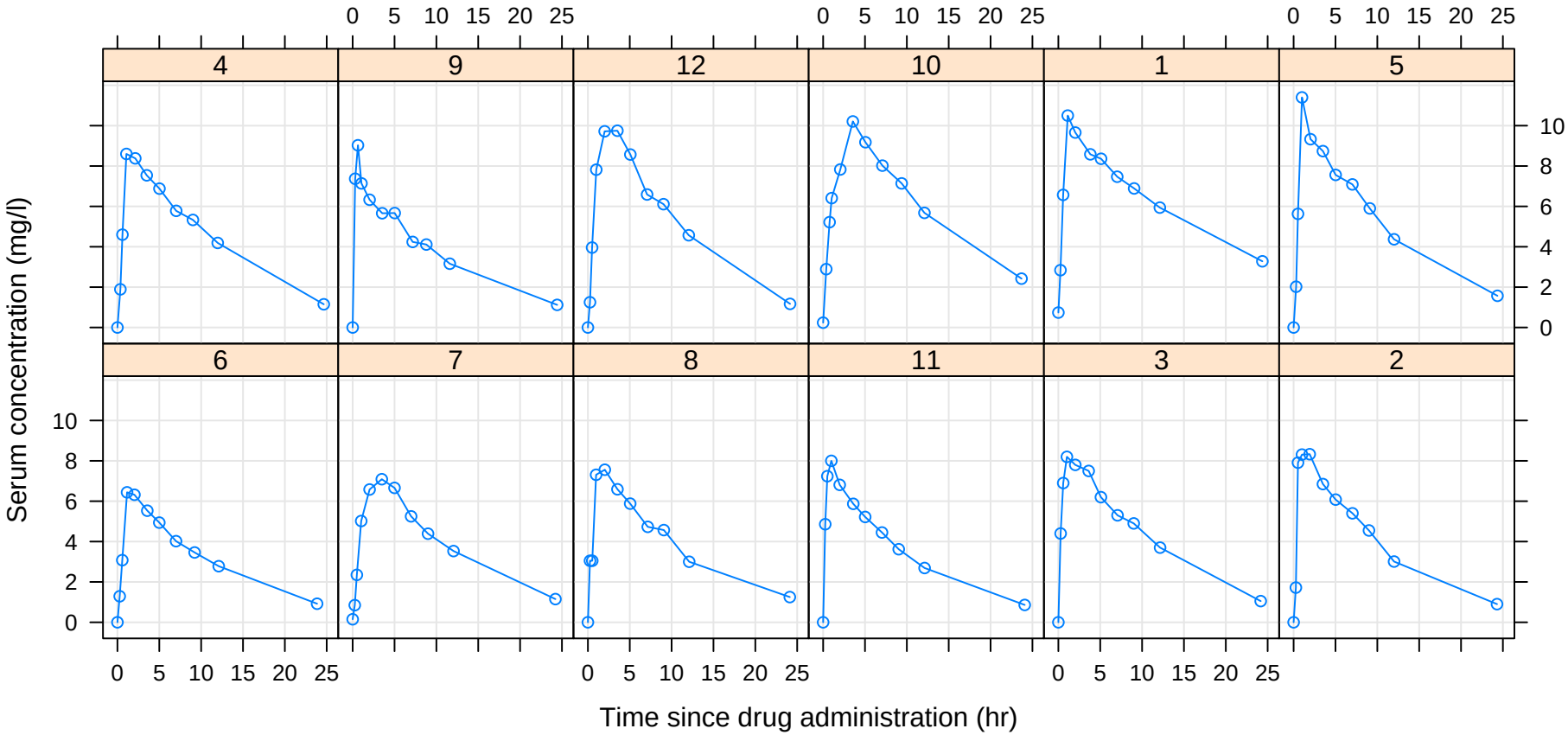
Outline

Outline

Introduction

- Population pharmacokinetic data are often modeled using *nonlinear mixed-effects models* (NLMMs).
- These are *nonlinear* because pharmacokinetic parameters - rate constants, clearance rates, etc. - occur nonlinearly in the model function.
- In statistical terms these are *mixed-effects models* because they involve both *fixed-effects parameters*, applying to the entire population or well-defined subsets of the population, and *random effects* associated with particular experimental or observational units under study.
- Many algorithms for obtaining parameter estimates, usually "something like" the *maximum likelihood estimates* (MLEs), for such models have been proposed and implemented.
- Comparing different algorithms is not easy. Even understanding the definition of the model and the proposed algorithm is not easy.

An example: Theophylline pharmacokinetics



- These are serum concentration profiles for 12 volunteers after injestion of an oral dose of Theophylline, as described in Pinheiro and Bates (2000).

Modeling pharmacokinetic data with a nonlinear model

- These are longitudinal repeated measures data.
- For such data the time pattern of an individual's response is determined by pharmacokinetic parameters (e.g. rate constants) that occur nonlinearly in the expression for the expected response.
- The form of the nonlinear model is determined by the pharmacokinetic theory, not derived from the data.

$$d \cdot k_e \cdot k_a \cdot C \frac{e^{-k_e t} - e^{-k_a t}}{k_a - k_e}$$

- These pharmacokinetic parameters vary over the population. We wish to characterize typical values in the population and the extent of the variation.
- Thus, we associate random effects with the parameters, k_a, k_e and C in the nonlinear model.

Outline

Statistical theory and applications - why we need both

- For 30 years, I have had the pleasure of being part of the U. of Wisconsin-Madison Statistics Dept. This year we celebrate the 50th anniversary of the founding of our department by George Box (who turned 90 earlier this year).
- George's approach, emphasizing **both** the theory and the applications of statistics, has now become second-nature to me.
- We are familiar with the dangers of practicing theory without knowledge of applications. As George famously said, "All models are wrong; some models are useful." How can you expect to decide if a model is useful unless you use it?
- We should equally be wary of the application of statistical techniques for which we know the "how" but not the "why". Despite the impression we sometimes give in courses, applied statistics is not just a "black box" collection of formulas into which you pour your data, hoping to get back a p-value that is less than 5%. (In the past many people felt that "applied statistics is the use of SAS" but now we know better.)

The evolving role of approximation

- When Don Watts and I wrote a book on nonlinear regression we included a quote from Bertrand Russell, "Paradoxically, all exact science is dominated by the idea of approximation". In translating statistical theory to applied techniques (computing algorithms) we almost always use some approximations.
- Sometimes the theory is deceptively simple (maximum likelihood estimates are the values of the parameters that maximize the likelihood, given the data) but the devil is in the details (so exactly how do I maximize this likelihood?).
- Decades of work by many talented people have provided us with a rich assortment of computational approximations and other tricks to help us get to the desired answer - or at least close to the desired answer.
- It is important to realize that approximations, like all aspects of computing, have a very short shelf life. Books on theory can be useful for decades; books on computing may be outmoded in a few years.

Failure to revisit assumptions leads to absurdities

- Forty years ago, when I took an intro engineering stats class, we used slide rules or pencil and paper for calculations. Our text took this into account, providing short-cut computational formulas and "rules of thumb" for the use of approximations, plus dozens of pages of tables of probabilities and quantiles.
- Today's computing resources are unimaginably more sophisticated yet the table of contents of most introductory text hasn't changed.
- The curriculum still includes using tables to evaluate probabilities, calculating coefficient estimates of a simple linear regression by hand, creating histograms (by hand, probably) to assess a density, approximating a binomial by a Poisson or by a Gaussian for cases not available in the tables, etc.
- Then we make up PDF slides of this content and put the file on a web site for the students to download and follow on their laptops during the lecture. Apparently using the computer to evaluate the probabilities or to fit a model would be cheating - you are supposed to do this by hand.

And what about nonlinear mixed-effects models?

- Defining the statistical model is subtle and all methods proposed for determining parameter estimates use approximations.
- Often the many forms of approximations are presented as different "types" of estimates from which one can pick and choose.
- In 2007-2008 a consortium of pharma companies, the NLMEc, discussed "next generation" simulation and estimation software for population PK/PD modeling. They issued a set of user requirements for such software including, in section 4.4 on estimation

  > *The system will support but not be limited to the following estimation methods: FO, FOI, FOCE, FOCEI, Laplacian, Lindstrom and Bates, MCMC, MCPEM, SAEM, Gaussian quadrature, and nonparametric methods.*

- Note the emphasis on estimation methods (i.e. algorithms). All of these techniques are supposed to approximate the mle's but that is never mentioned.

Outline

Linear and nonlinear mixed-effects models

- Both linear and nonlinear mixed-effects models, are based on the n-dimensional response random variable, $\boldsymbol{\mathcal{Y}}$, whose value, $\boldsymbol{y}$, is observed, and the q-dimensional, unobserved random effects variable, $\boldsymbol{\mathcal{B}}$.
- In the models we will consider $\boldsymbol{\mathcal{B}} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_\theta)$. The variance-covariance matrix $\boldsymbol{\Sigma}_\theta$ can be huge but it is completely determined by a small number of *variance-component parameters*, $\boldsymbol{\theta}$.
- The conditional distribution of the response, $\boldsymbol{\mathcal{Y}}$, is

$$(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}} = \boldsymbol{b}) \sim \mathcal{N}\left(\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}}, \sigma^2 \boldsymbol{I}_n\right)$$

- The conditional mean, $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}}$, depends on $\boldsymbol{b}$ and on the fixed-effects parameters, $\boldsymbol{\beta}$, through a *linear predictor* expression, $\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}\boldsymbol{\beta}$.
- For a linear mixed model (LMM), $\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{B}}}$ is exactly the linear predictor. For an NLMM the linear predictor determines the parameter values in the nonlinear model function which then determines the mean.

Transforming to orthogonal random effects

- We never really form $\boldsymbol{\Sigma}_\theta$; we always work with the *relative covariance factor*, $\boldsymbol{\Lambda}_\theta$, defined so that

$$\boldsymbol{\Sigma}_\theta = \sigma^2 \boldsymbol{\Lambda}_\theta \boldsymbol{\Lambda}_\theta^\mathsf{T}.$$

  Note that we must allow for $\boldsymbol{\Lambda}_\theta$ to be less that full rank.

- We define a q-dimensional "spherical" or "unit" random-effects vector, $\mathcal{U}$, such that

$$\mathcal{U} \sim \mathcal{N}\left(\mathbf{0}, \sigma^2 \boldsymbol{I}_q\right), \ \mathcal{B} = \boldsymbol{\Lambda}_\theta \mathcal{U} \Rightarrow \mathsf{Var}(\mathcal{B}) = \sigma^2 \boldsymbol{\Lambda}_\theta \boldsymbol{\Lambda}_\theta^\mathsf{T} = \boldsymbol{\Sigma}_\theta.$$

- Setting $\boldsymbol{U}_\theta = \boldsymbol{Z}\boldsymbol{\Lambda}_\theta$, the linear predictor expression becomes

$$\boldsymbol{Z}\boldsymbol{b} + \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{Z}\boldsymbol{\Lambda}_\theta \boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{U}_\theta \boldsymbol{u} + \boldsymbol{X}\boldsymbol{\beta}.$$

  where $\boldsymbol{U}_\theta$, like $\boldsymbol{Z}_\theta$ is a large, sparse matrix.

The conditional mode, $\tilde{\boldsymbol{u}}_{\theta,\beta}$

- Although the probability model is defined from $(\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}=\boldsymbol{u})$, we observe $\boldsymbol{y}$, not $\boldsymbol{u}$ (or $\boldsymbol{b}$) so we want to work with the other conditional distribution, $(\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y})$.
- The joint distribution of $\boldsymbol{\mathcal{Y}}$ and $\boldsymbol{\mathcal{U}}$ is Gaussian with density

$$\begin{aligned} f_{\boldsymbol{\mathcal{Y}},\boldsymbol{\mathcal{U}}}(\boldsymbol{y},\boldsymbol{u}) &= f_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}(\boldsymbol{y}|\boldsymbol{u})\, f_{\boldsymbol{\mathcal{U}}}(\boldsymbol{u}) \\ &= \frac{\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}\|^2)}{(2\pi\sigma^2)^{n/2}}\,\frac{\exp(-\frac{1}{2\sigma^2}\|\boldsymbol{u}\|^2)}{(2\pi\sigma^2)^{q/2}} \\ &= \frac{\exp(-\left[\|\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}\|^2+\|\boldsymbol{u}\|^2\right]/(2\sigma^2))}{(2\pi\sigma^2)^{(n+q)/2}} \end{aligned}$$

- The mode, $\tilde{\boldsymbol{u}}_{\theta,\beta}$, of the conditional distribution $(\boldsymbol{\mathcal{U}}|\boldsymbol{\mathcal{Y}}=\boldsymbol{y})$ (also the conditional mean in the case of an LMM) is

$$\tilde{\boldsymbol{u}}_{\theta,\beta} = \arg\min_{\boldsymbol{u}} \left[\left\|\boldsymbol{y}-\boldsymbol{\mu}_{\boldsymbol{\mathcal{Y}}|\boldsymbol{\mathcal{U}}}\right\|^2+\|\boldsymbol{u}\|^2\right]$$

Outline

Minimizing a penalized sum of squared residuals

- An expression like $\left\|\boldsymbol{y}-\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}\right\|^2+\|\boldsymbol{u}\|^2$ is called a *penalized sum of squared residuals* because $\left\|\boldsymbol{y}-\boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}}\right\|^2$ is a sum of squared residuals and $\|\boldsymbol{u}\|^2$ is a penalty on the size of the vector $\boldsymbol{u}$.
- Determining $\tilde{\boldsymbol{u}}_{\theta,\beta}$ as the minimizer of this expression is a *penalized least squares* (PLS) problem. For an LMM it is a *penalized linear least squares problem* that can be solved directly (i.e. without iterating). For an NLMM it is a *penalized nonlinear least squares problem*.
- One way to determine the solution in an LMM is to rephrase it as a linear least squares problem for an extended residual vector

$$\tilde{\boldsymbol{u}}_{\theta,\beta}=\arg\min_{\boldsymbol{u}}\left\|\begin{bmatrix}\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\\ \boldsymbol{0}\end{bmatrix}-\begin{bmatrix}\boldsymbol{U}_\theta\\ \boldsymbol{I}_q\end{bmatrix}\boldsymbol{u}\right\|^2$$

This is sometimes called a *pseudo-data* approach because we create the effect of the penalty term, $\|\boldsymbol{u}\|^2$, by adding "pseudo-observations" to $\boldsymbol{y}$ and to the predictor.

The profiled deviance for LMMs

- We can see that $\tilde{\boldsymbol{u}}_{\theta,\beta}$ satisfies $\left(\boldsymbol{U}_\theta^\mathsf{T}\boldsymbol{U}_\theta + \boldsymbol{I}_q\right)\tilde{\boldsymbol{u}}_{\theta,\beta} = \boldsymbol{U}_\theta^\mathsf{T}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ which we solve using the sparse Cholesky decomposition

$$\boldsymbol{L}_\theta\boldsymbol{L}_\theta^\mathsf{T} = \boldsymbol{P}\left(\boldsymbol{U}_\theta^\mathsf{T}\boldsymbol{U}_\theta + \boldsymbol{I}_q\right)\boldsymbol{P}^\mathsf{T}$$

  $\boldsymbol{P}$ is a permutation matrix that has practical importance but does not affect the theory. The matrix $\boldsymbol{L}_\theta$ is the sparse, lower-triangular factor.
- Let $r^2(\boldsymbol{\theta},\boldsymbol{\beta})$ be the minimum penalized residual sum of squares, then $\ell(\boldsymbol{\theta},\boldsymbol{\beta},\sigma|\boldsymbol{y}) = \log L(\boldsymbol{\theta},\boldsymbol{\beta},\sigma|\boldsymbol{y})$ can be written

$$-2\ell(\boldsymbol{\theta},\boldsymbol{\beta},\sigma|\boldsymbol{y}) = n\log(2\pi\sigma^2) + \frac{r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{\sigma^2} + \log(|\boldsymbol{L}_\theta|^2)$$

- The conditional estimate of σ^2 is

$$\widehat{\sigma^2}(\boldsymbol{\theta},\boldsymbol{\beta}) = \frac{r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{n}$$

  producing the *profiled deviance*

$$-2\tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\beta}|\boldsymbol{y}) = \log(|\boldsymbol{L}_\theta|^2) + n\left[1 + \log\left(\frac{2\pi r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{n}\right)\right]$$

Profiling the deviance with respect to $\boldsymbol{\beta}$ for LMMs

- In a LMM the deviance depends on $\boldsymbol{\beta}$ only through $r^2(\boldsymbol{\theta},\boldsymbol{\beta})$ we can obtain the conditional estimate, $\widehat{\boldsymbol{\beta}}_\theta$, by extending the PLS problem to

$$r^2(\boldsymbol{\theta}) = \min_{\boldsymbol{u},\boldsymbol{\beta}} \left[\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{U}_\theta\,\boldsymbol{u}\|^2 + \|\boldsymbol{u}\|^2 \right]$$

  with the solution satisfying the equations

$$\begin{bmatrix} \boldsymbol{U}_\theta^\mathsf{T}\boldsymbol{U}_\theta + \boldsymbol{I}_q & \boldsymbol{U}_\theta^\mathsf{T}\boldsymbol{X} \\ \boldsymbol{X}^\mathsf{T}\boldsymbol{U}_\theta & \boldsymbol{X}^\mathsf{T}\boldsymbol{X} \end{bmatrix} \begin{bmatrix} \tilde{\boldsymbol{u}}_\theta \\ \widehat{\boldsymbol{\beta}}_\theta \end{bmatrix} = \begin{bmatrix} \boldsymbol{U}_\theta^\mathsf{T}\boldsymbol{y} \\ \boldsymbol{X}^\mathsf{T}\boldsymbol{y}. \end{bmatrix}$$

- The profiled deviance, which is a function of $\boldsymbol{\theta}$ only, is

$$-2\tilde{\ell}(\boldsymbol{\theta}) = \log(|\boldsymbol{L}_\theta|^2) + n\left[1 + \log\left(\frac{2\pi r^2(\boldsymbol{\theta})}{n}\right)\right]$$

Conditional mode and profiled Laplace approximation for NLMMs

- As previously stated, determining the conditional mode

$$\tilde{\boldsymbol{u}}_{\theta,\beta} = \arg\min_{\boldsymbol{u}} \left[\left\| \boldsymbol{y} - \boldsymbol{\mu}_{\mathcal{Y}|\mathcal{U}} \right\|^2 + \left\| \boldsymbol{u} \right\|^2 \right]$$

  in an NLMM is a penalized nonlinear least squares (PNLS) problem.
- It is a nonlinear optimization problem but a comparatively simple one. The penalty term *regularizes* the optimization.
- The Laplace approximation to the profiled deviance (profiled over σ^2) is, as before,

$$-2\tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\beta}|\boldsymbol{y}) = \log(|\boldsymbol{L}_\theta|^2) + n\left[1 + \log\left(\frac{2\pi r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{n}\right)\right]$$

  where $\boldsymbol{L}_\theta$ is the sparse Cholesky factor evaluated at the conditional mode.
- The motivation for this approximation is that it replaces the conditional distribution, $(\mathcal{U}|\mathcal{Y}=\boldsymbol{y})$, for parameters $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and σ, by a multivariate Gaussian approximation, *evaluated at the mode*.

Laplace approximation and adaptive Gauss-Hermite quadrature

- The Laplace approximation

$$-2\tilde{\ell}(\boldsymbol{\theta},\boldsymbol{\beta}|\boldsymbol{y}) = \log(|\boldsymbol{L}_\theta|^2) + n\left[1 + \log\left(\frac{2\pi r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{n}\right)\right]$$

  is a type of *smoothing objective* consisting of two terms: $n\left[1 + \log\left(\frac{2\pi r^2(\boldsymbol{\theta},\boldsymbol{\beta})}{n}\right)\right]$, which measures fidelity to the data, and $\log(|\boldsymbol{L}_\theta|^2)$, which measures the complexity of the model.
- For models with a simple structure for the random effects (the matrices $\boldsymbol{\Sigma}_\theta$ and $\boldsymbol{\Lambda}_\theta$ are block diagonal consisting of a large number of small blocks) a further enhancement is to use *adaptive Gauss-Hermite quadrature*, requiring values of the RSS at several points near $\tilde{\boldsymbol{u}}_{\theta,\beta}$
- Note that the modifier *adaptive*, meaning evaluating at the conditional mode, is important. Gauss-Hermite quadrature without first determining the conditional mode is not a good idea.

Outline

Consequences for comparisons of methods

- We should distinguish between an algorithm, which is a sort of a black box, and a criterion, such as maximizing the likelihood (or, equivalently, minimizing the deviance.
- The criterion is based on the statistical model and exists outside of any particular implementation or computing hardware. It is part of the theory, which has a long shelf life.
- A particular approximation, algorithm and implementation has a short shelf life.
- I claim it does not make sense to regard the FO, FOI, ... methods as producing well-defined types of "estimates" in the same sense that maximum likelihood estimates, or maximum *a posteriori* estimates are defined.
- If you use a criterion to define an estimation method then implementations should be compared on the basis of that criterion, not on something like mean squared error.