An overview of non-parametric estimation methods used in population analysis

Robert H. Leary

Fellow, Certara, Cary, N.C.

Goals of presentation - answer

What does a non-parametric population pk/pd method do that I can't do with a parametric method?

What does a NP estimate look like?

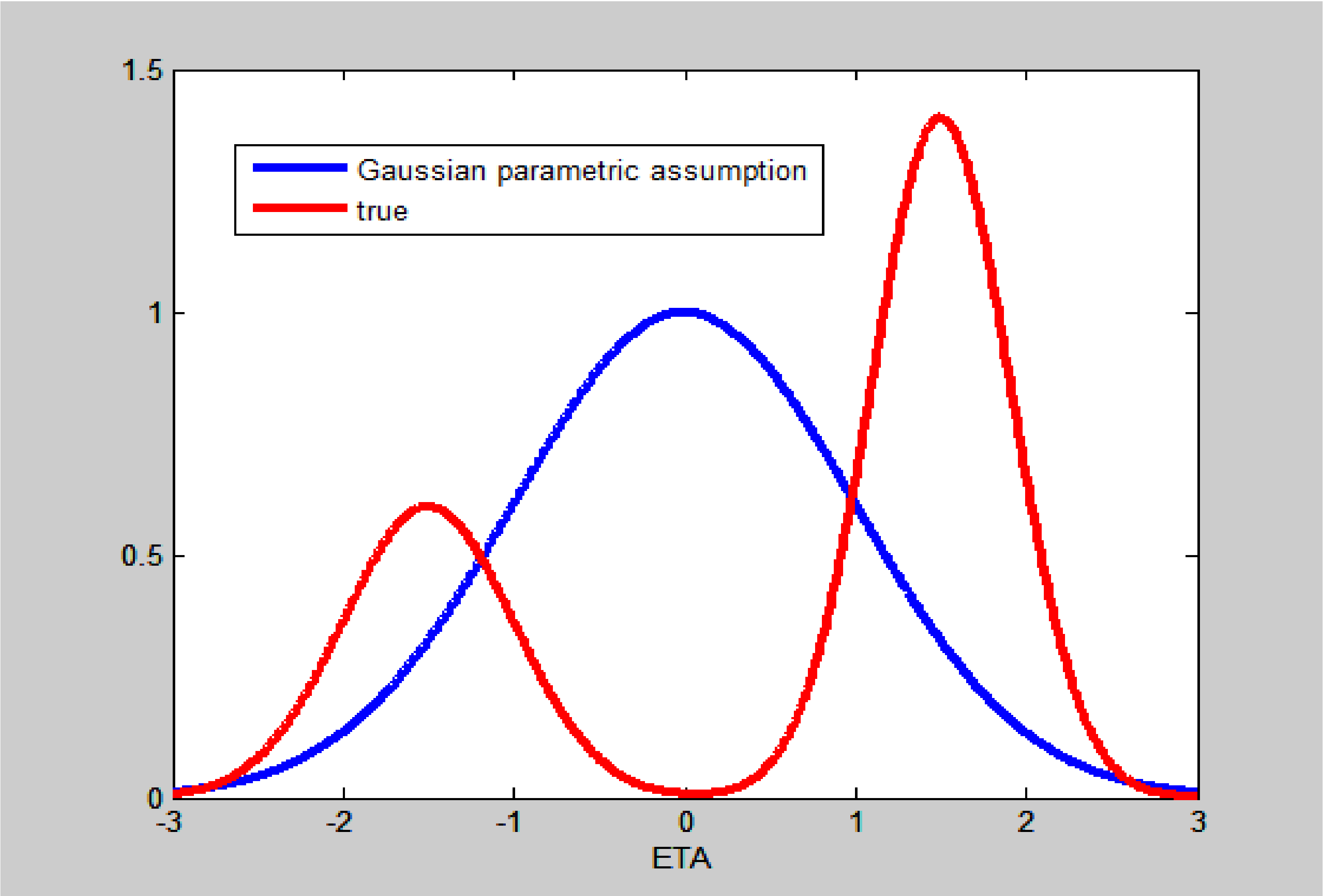
Where can I get access to an NP estimation method, and how do the commonly available ones differ?

What are the advantages and disadvantages of NP estimation?

What's new in the field?

What still needs to be done?

Focus of non-parametric methods is on random effects



Some estimation approaches for the non-normal case

Mixture distributions ($MIX in NM, mixture of normals)

Non-parametric maximum likelihood

Non-parametric Bayesian (Dirichlet stick breaking priors)

Smooth non-parametric (e.g., polynomial times normal)

Semi-parametric maximum likelihood
(NP for random effects,
some parametric components as
in residual error function or fixed
effects in covariate models)

A brief history of likelihood-based NP approaches

1983 – B. Lindsay - maximum likelihood non- parametric estimator is a discrete distribution with at most Nsub support points

1986 – A. Mallet – NPML using vertex direction method; established connection to optimal design criterion and use of directional derivative

1992 – A. Schumitzky – NPEM – EM algorithm on a single large fixed grid of support points

1998 – present , Laboratory of Applied Pharmacokinetics– NPAG, NPOD with primal dual probability optimization and mobile support points resulting in greatly increased speed and accuracy relative to NPEM, as well as complementary multiple model dosage program (currently called Pmetrics and BestDose – NPAG variant is also in Phoenix NLME)

2006 – present - NONMEM NONP method based on using EBE's from a prior parametric method run as support points, EM for probability optimization, and residual error and covariate model fixed effects from parametric. Extended by M. Karlsson's group with 'extended grid' methodology, now available in latest 2017 NONMEM release.

Some recent Ph.D. theses

X. Wang, 2015, U. of Auckland, “Mixture Models for Multivariate Observations”

P. Baveral, 2011, Uppsala, “Development and Evaluation of Nnonparametric Mixed Effects Models

J. Antic, 2009, Université Paul Sabatier – Toulouse, “Nonparametric methods for population PK and/or PD”

Y. Baek, 2006, Univ. of Washington, “An Interior Point Approach to Constrained Nonparametric Mixture Models”

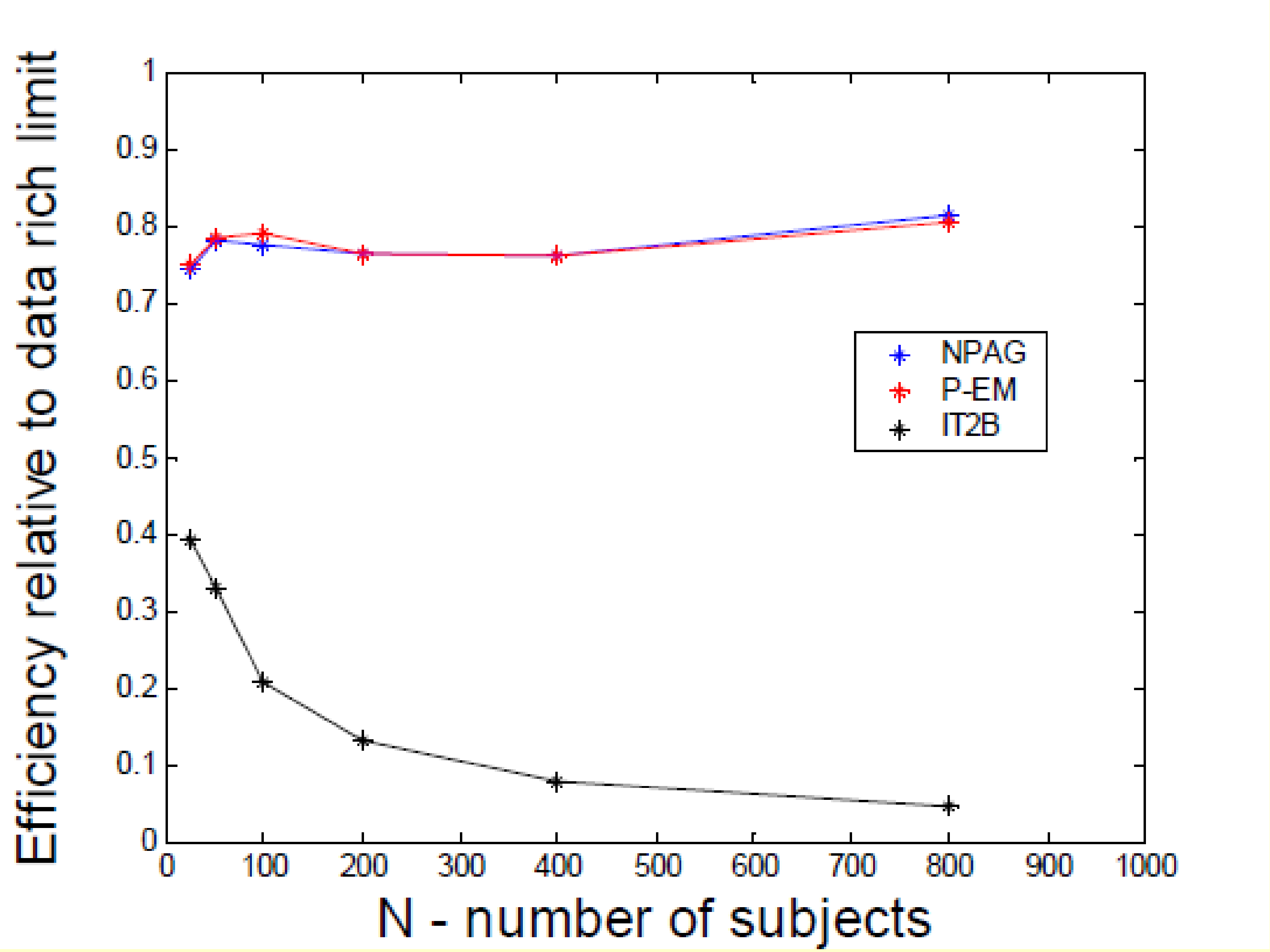
Quick review of ML estimation properties

Consistent (true in parametric and non-parametric case) (in parametric case, parameters converge to true values as amount of data becomes large; in NP cumulative distribution function converges to true cumulative distribution function).

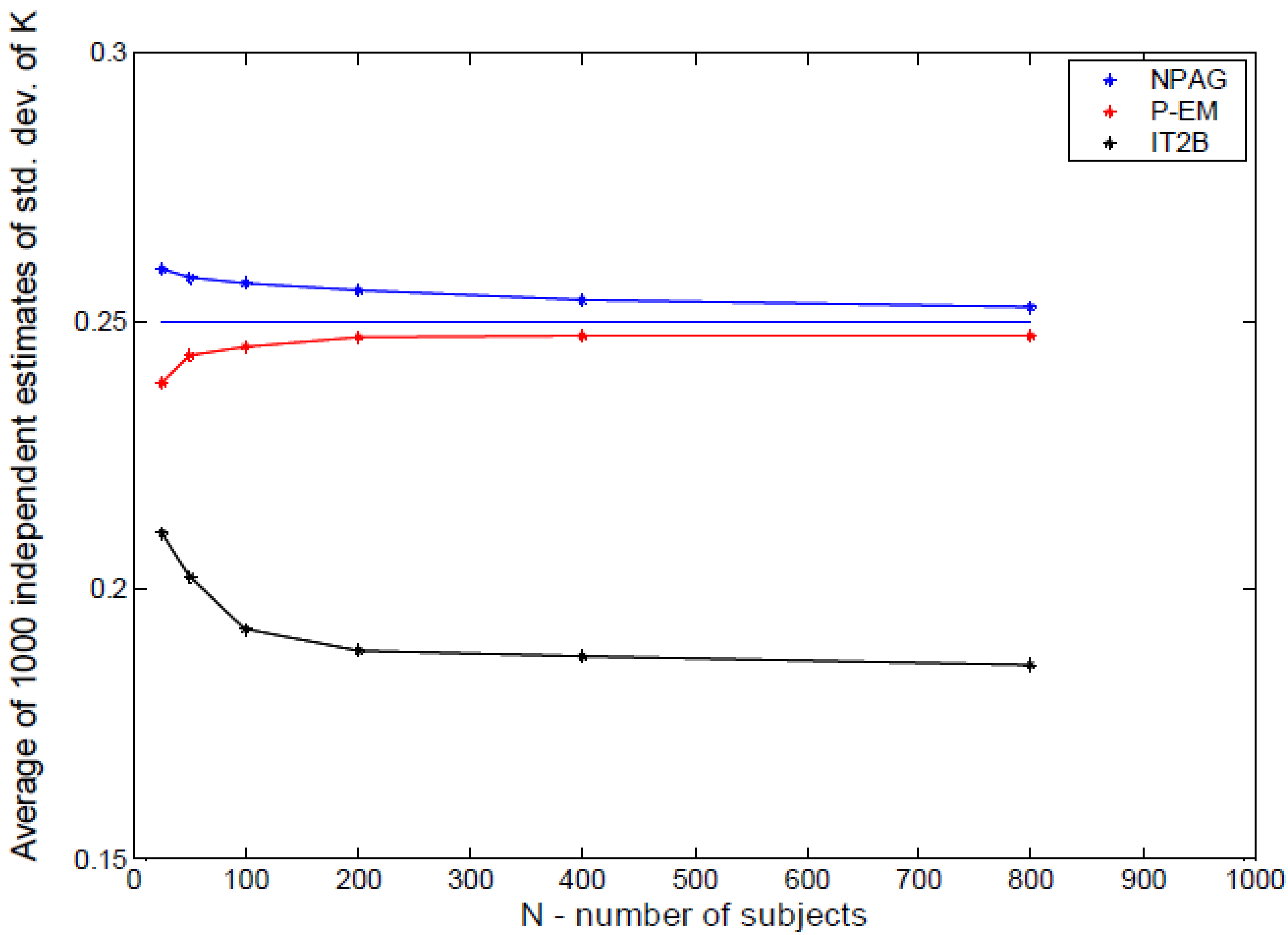
Asymptotically most efficient estimator in parametric case, with asymptotically normal parameter estimates. Leads to Fisher information based estimates of parameter uncertainty. N/A in NP case, so some other methodology such as bootstrapping is necessary.

From Leary et al, PAGE 2002

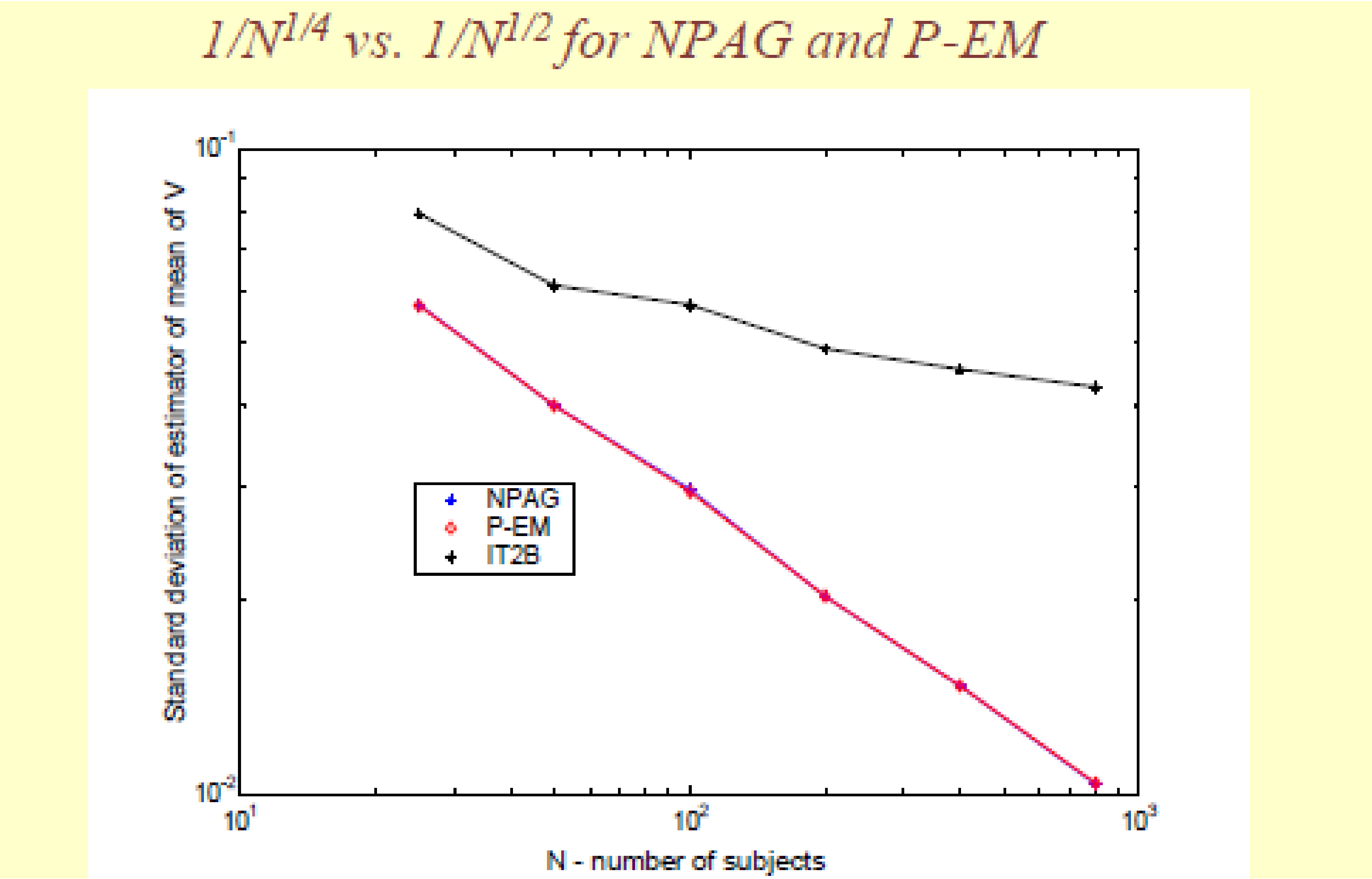
Statistical efficiencies of NPAG and P-EM are nearly identical and much higher than IT2B



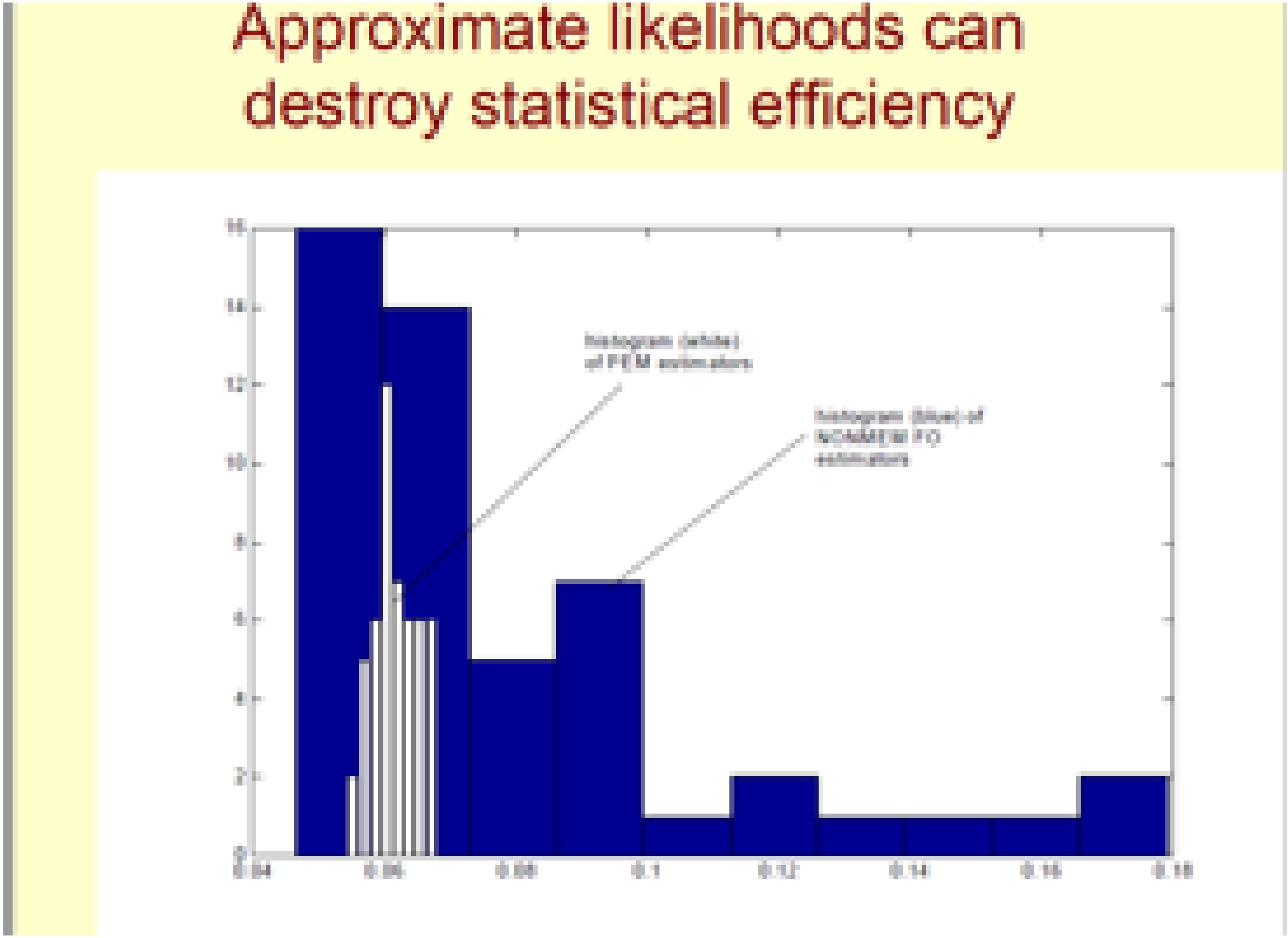
Consistency of estimators of σ_K (true value of $\sigma_K=0.25$)



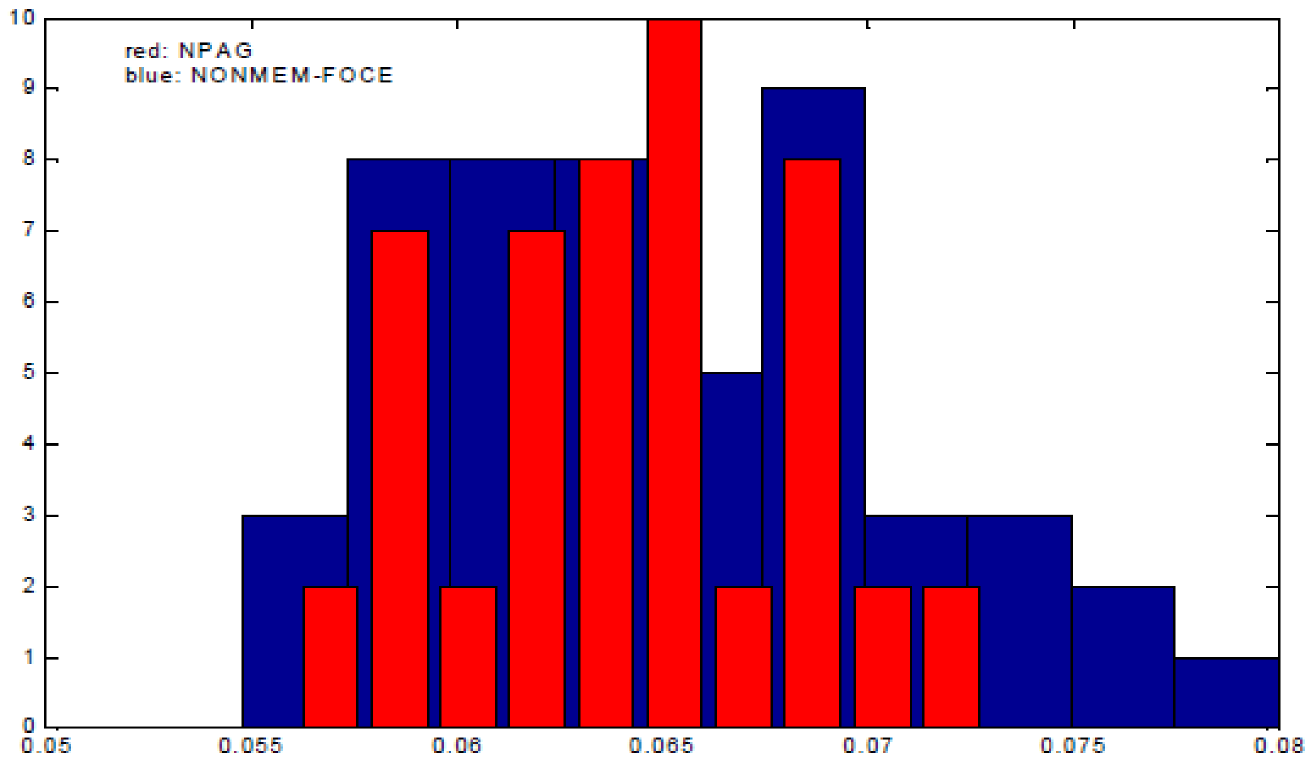
Asymptotic convergence rate of NPAG is same as parametric EM



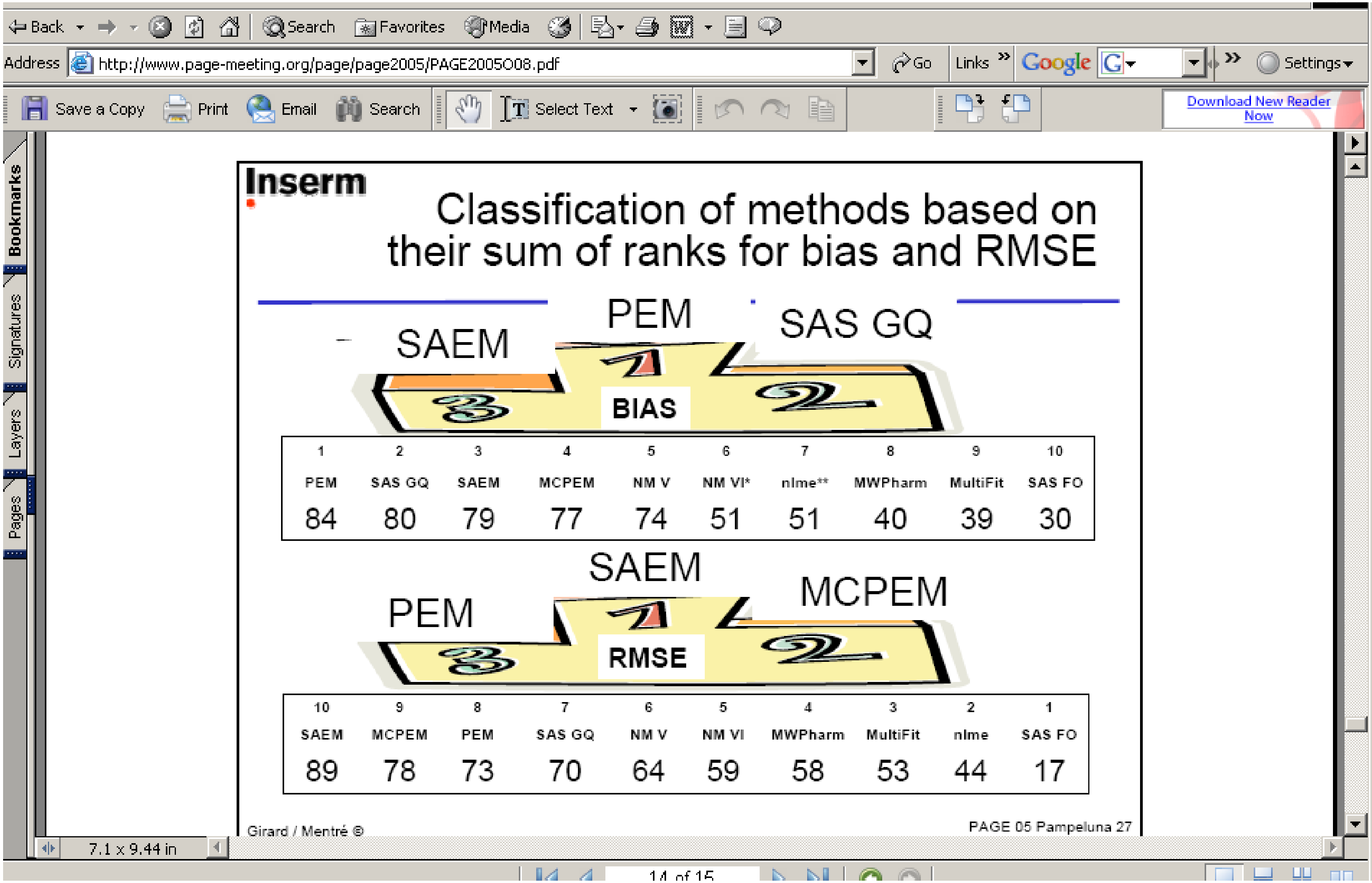
From Leary et al, PAGE2004



FOCE does better, but still has <40% efficiency relative to ML



Results from 2004-2005 inter-method blind comparison



Classification of methods based on their sum of ranks for bias and RMSE

1	2	3	4	5	6	7	8	9	10
PEM	SAS GQ	SAEM	MCPEM	NM V	NM VI*	nlme**	MWPharm	MultiFit	SAS FO
84	80	79	77	74	51	51	40	39	30

10	9	8	7	6	5	4	3	2	1
SAEM	MCPEM	PEM	SAS GQ	NM V	NM VI	MWPharm	MultiFit	nlme	SAS FO
89	78	73	70	64	59	58	53	44	17

Girard / Mentré ©

PAGE 05 Pampeluna 27

Non-parametric log likelihood LL(f(ETA)) is concave

$$LL(f) = \sum_{isub=1}^{Nsub} \log\left(\int l\_isub(res|ETA)\, f(ETA) dETA\right)$$

for f_1, f_2 *probability densities,*

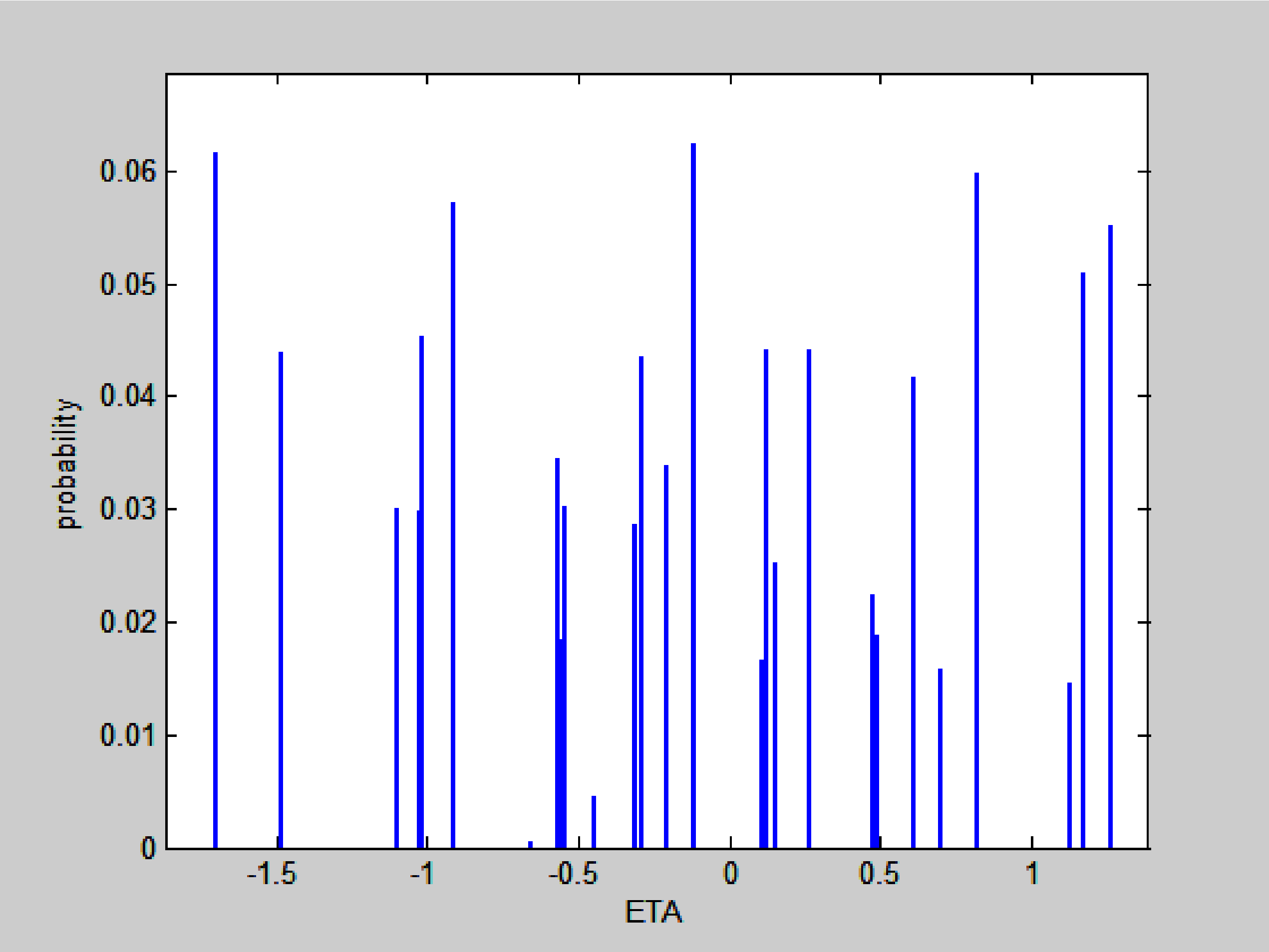
$f = (1-\alpha) f_1 + \alpha f_2, 0 \leq \alpha \leq 1$ is a probability density and

$$LL(f) \geq (1-\alpha) LL(f_1) + \alpha (LL(f_2))$$

If f is discrete on a grid of Npoint support points

$$LL(f) = \sum_{isub=1}^{Nsub} \log\left(\sum_{jpoint=1}^{Npoint} L(isub, ETA(jpoint)) * p(jpoint)\right)$$

Theorem – NPML estimate is discrete with at most Nsub points



Two parts to maximizing the NP likelihood

1. Find optimal grid G={ETA(jpoint), jpoint=1,Npoint}

2. Find optimal probabilities p(jpoint), jpoint=1,Npoint

Current and historical methods differ by how they approach these two parts and whether they iterate over both parts

NPML iterates over both 1 and 2, but does only an approximate job of getting optimal probabilities at each iteration, and only changes adds one new grid point per iteration

NPEM starts with a large fixed grid, does only 1 iteration but with accurate probabilities found by EM on the intial grid

NONP starts with a relatively small initial grid of 'good' points, does one accurate probability optimization using EM, but like NPEM does not change the grid. 2017 NONMEM7.4 now supports Uppsala 'extended grid' methodology

NPAG and NPOD iterate over relatively small grids, use a fast quadratically convergent primal dual probability optimization , change out multiple grid points per iteration, and do at least some form of grid point position selection

Improving the current grid with a new candidate point ETA

Let {G={ETA(j)}, p} be any grid and the associated optimized probabilities, and L(isub,jpoint) the corresponding likelihood matrix

Then a better grid can be found by adding any point ETA such that DD(ETANEW)>0, where DD(eta) is the directional derivative

defined by

DD(ETA) = w’*Likelihood(ETA))

where w’(isub) =1./(L*p)

The coeficients vector w is the ‘dual solution’

Suggests maximizing DD(ETA) over ETA to find new grid points (NPOD idea) or at least using w to screen new candidate grid points – (NPAG) - this essentially solves the ‘curse of dimensionality’ problem

Probability optimization methods

EM - used in NPEM, NONP
Reliable, but at best linearly convergent. Very simple to implement (5 lines of MATLAB). Convergence rate slows down as number of grid points increases and grid point spacing gets smaller. Fair restart properties if you add additional points to grid following an initial solution. Requires interior point start.

Primal-Dual – used in Pmetrics NPAG, Phoenix NLME NPAG
Reliable, quadratically convergent. Much faster than EM. Very insensitive to grid spacing. Fair restart capabilities on new grid points. Requires interior point start.

Non-Negative Least Squares –"New" method proposed by Y. Wang (2007, 2010). Finite step convergence on each iteration. Insensitive to grid spacing and often faster than primal-dual. Good restart capabilities – does not need an interior point start so you can start directly from a previous optimal solution on the old grid.

EM Method

EM method works by updating p with empirical Bayesian distributions (EBDs) induced by p -

pEBD(isub,jpoint)~ L(isub,jpoint)*p(jpoint)

pnew(jpoint) = mean_isub pEBD(isub,jpoint)

NP log likelihood is guaranteed to increase on each iteration by Jensen's inequality.   All likelihoods at each stage are exact (up to precision of model evaluation - no numerical derivatives or matrix factorizations anywhere that can fail.

Primal Dual Interior Point Method (Burke, Baek, 2001)

Works by simultaneously solving the primal problems
Maximize sum (log(L*p)), p≥0, sum(p(jpoint))=1

And the dual
maximize sum (w(isub))
L^T*w≤ones(npoint,1)

Inequalities are replaced by log barrier functions with a coefficient That is gradually reduced to zero. A Newton method is used to Solve the resulting Kuhn-Tucker equations at each coefficient value. The required first and second derivatives with respect to p and w These can be evaluated analytically and only involve the already computed L(isub,jpoint) values. There is one Cholesky factorization, but the PD method carefully controls the condition number of the matrix so this is very reliable.

Non-negative Least Squares (Y. Wang, 2007, 2010)

Do a second order Taylor expansion around current guess p_0

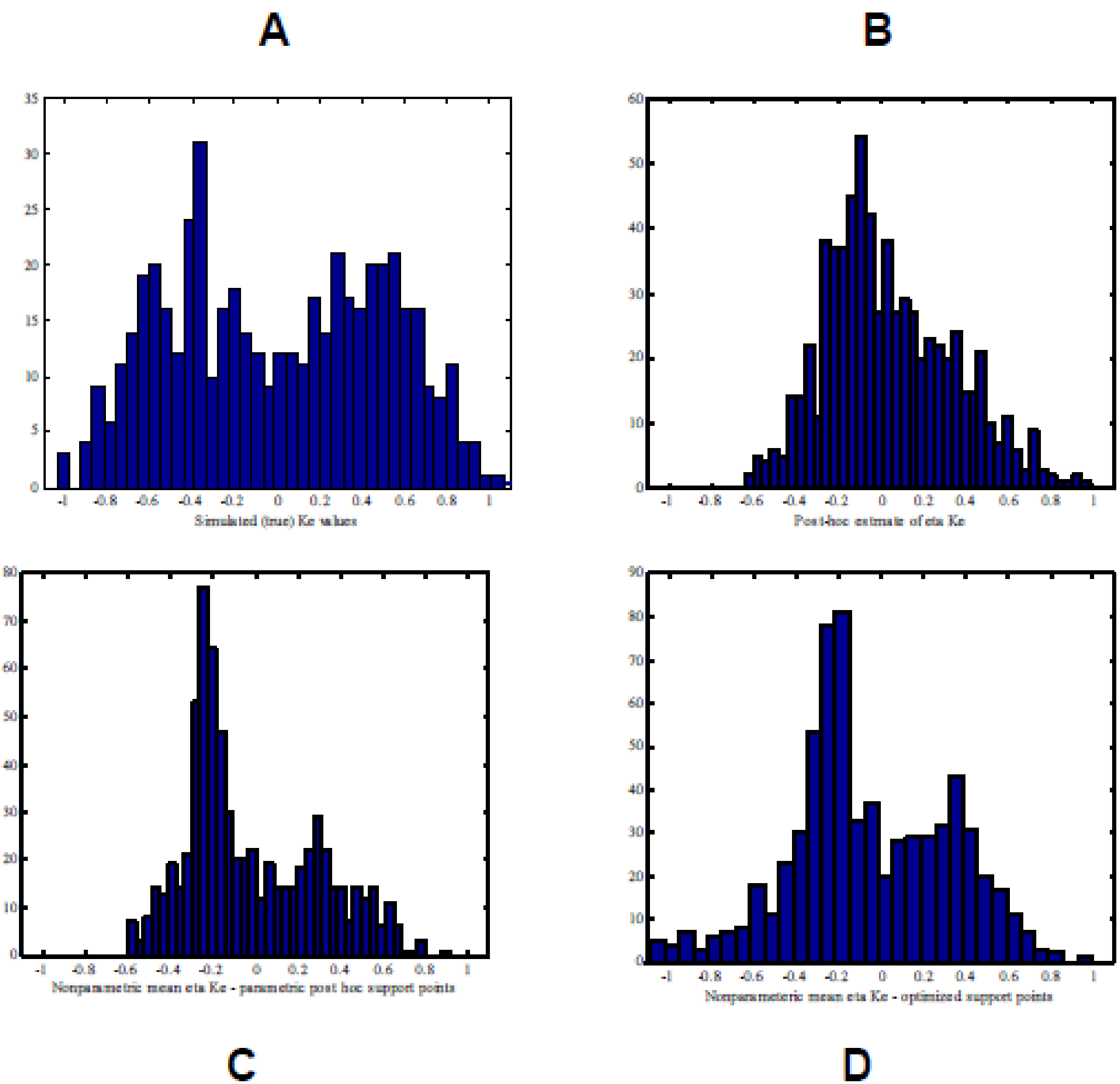
$LL(p) = LL\{p_0) + g^T(p-p_0) + (p-p_0)^T H(p-p_0)/2$

As in the primal dual method, all first and second derivatives are analytic and easily computed by simple linear algebra operations using L and p. Problem can be transformed easily to a non-negative least squares form

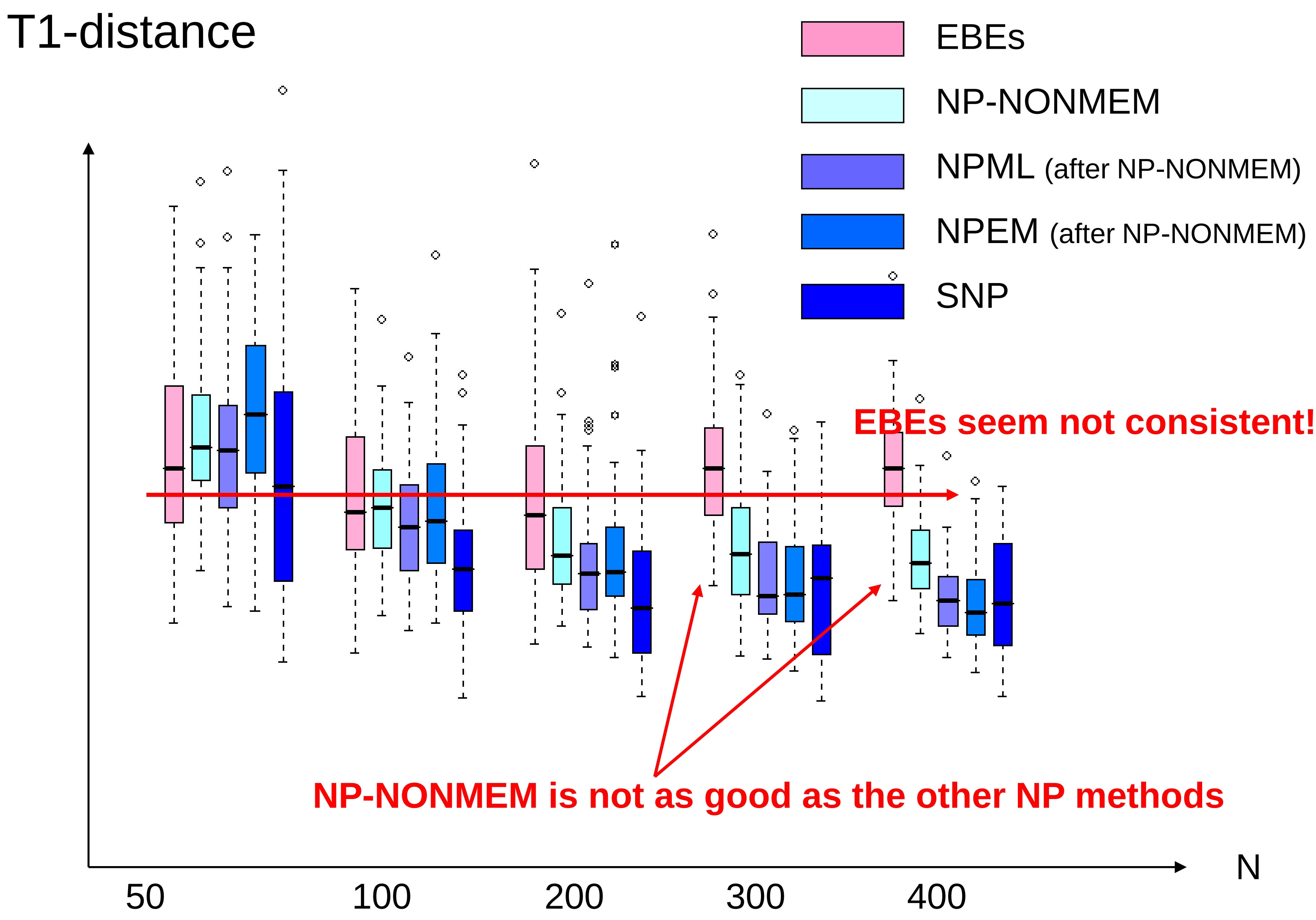
$$\underset{p \geq 0}{\text{minimize}} \; \|Ap-b\|^2$$

for which there are very fast algorithms that converge in a finite number of steps. Often there are fewer total floating point operations Than in the primal dual method and the final zero values in p are exact. May have to iterate once or twice on expanding in Taylor series step). This is very well suited to bootstrapping – each new replicate can often be solved in times <0.1 sec for 1000 subjects, 100 supports points.

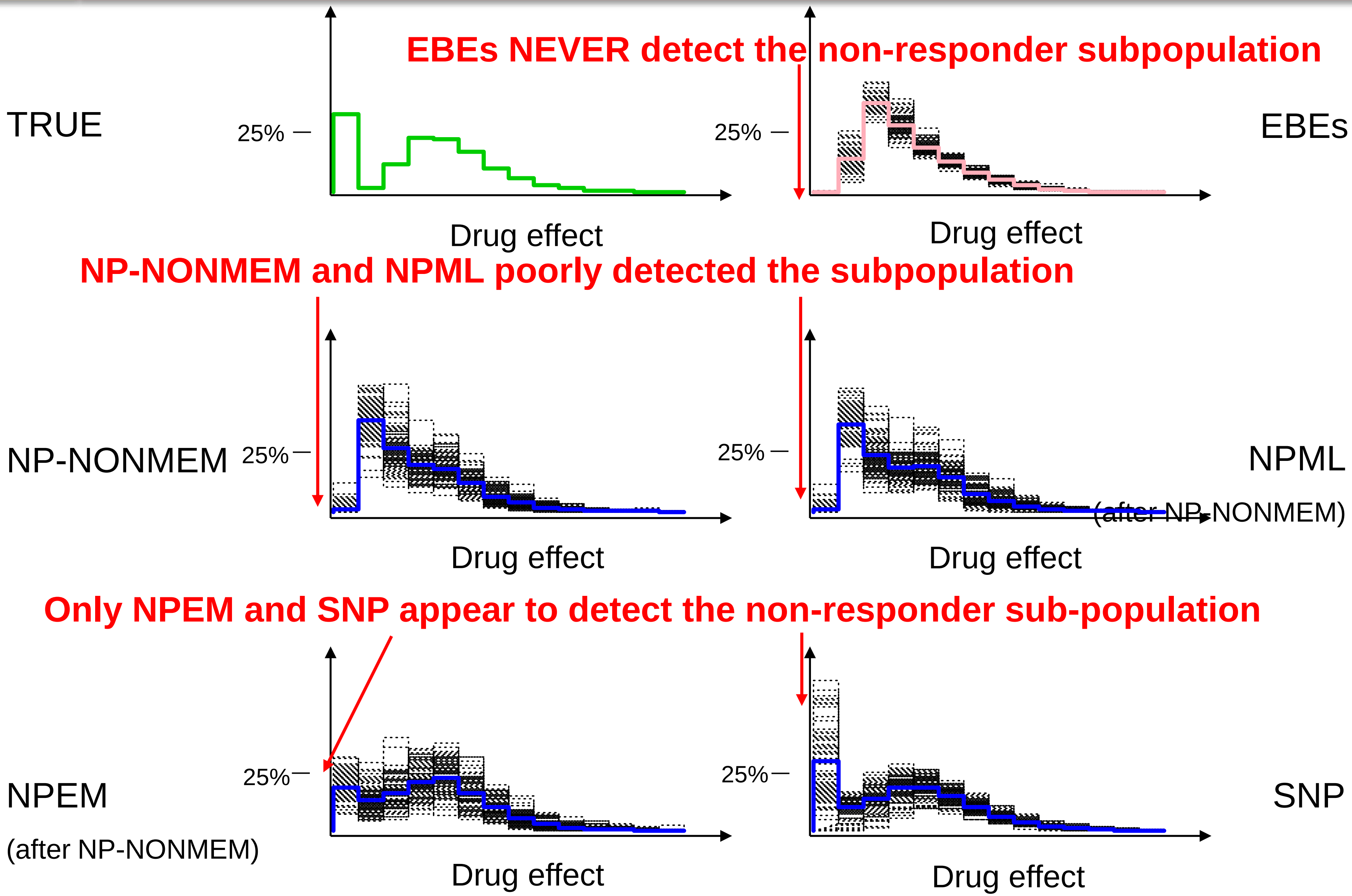
EBE’s as support points may not be good enough (Leary, PAGE 2007)



ETA-shrinkage ~ 31%; PK oral (from Antic, et al PAGE2009, abs. 1458)



ETA-shrinkage > 40%; PK/PD (from Antic, et al PAGE2009, abs. 1458)



Savic, Karlsson ,Extended Grid Method (AAPS Journal, V. 11, No. 3, 2009)

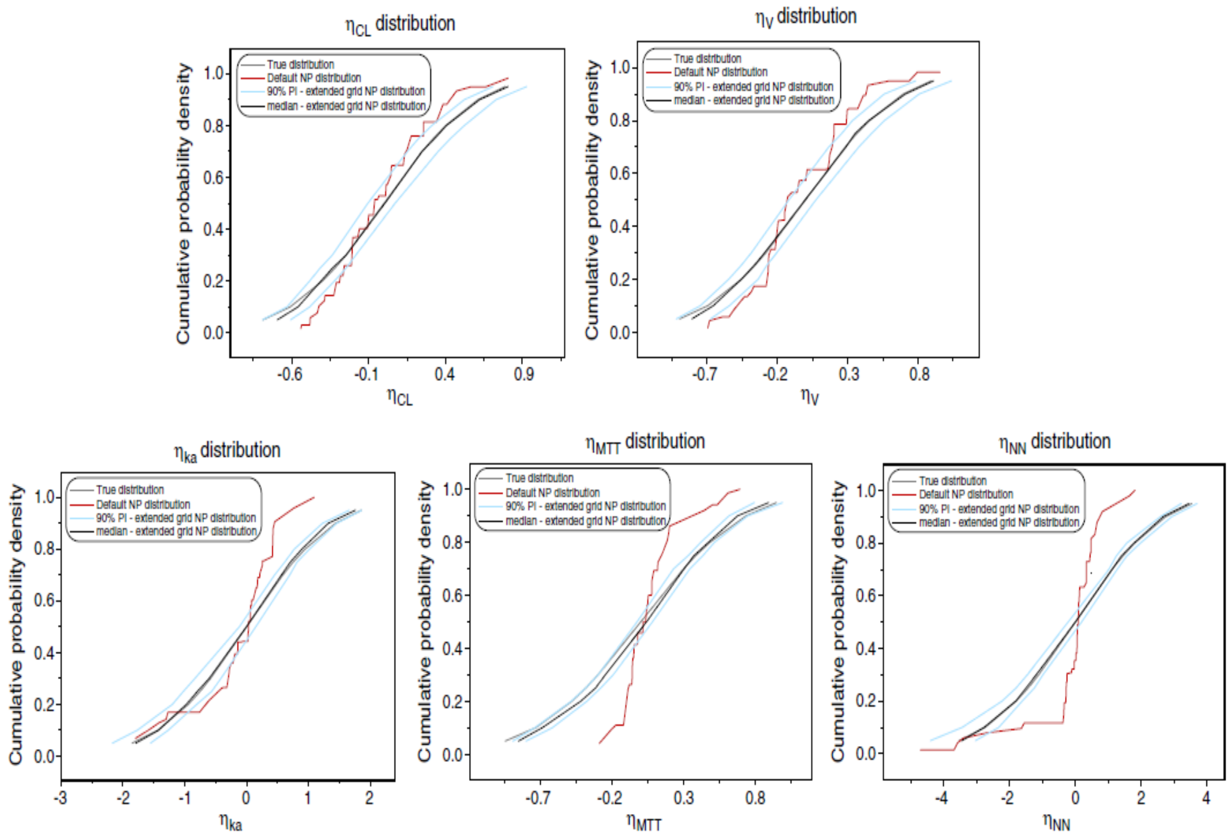


Fig. 6. Sparse pediatric example: 90% prediction intervals (*blue line*) and a median (*black line*) for each of the five parameter distributions created based on 100 estimated nonparametric distributions using extended grid method (1,000 additional support points) and its comparison with the true distribution (*gray line*) and default nonparametric distribution estimated with the original real data (*red line*)

Lessons learned

Parametric EBE’s are not good enough as a grid when shrinkage is large, but parametric solution is usually a good way to import residual error parameters and fixed effect values in covariate models

Better NP solutions require additional grid points

Directional derivative optimal design method works well in finding and filtering prospective new grid points

Modern optimization methods (primal dual, sequential quadratic programming) make probability optimization very fast – rate limiting step is often evaluation of model on each grid point.

Multiple Model Control

- **Discrete form of NP distribution is a natural fit for multiple model control in clinical dosing applications –e.g. *BestDose* software from LAPK. Rather than controlling a single profile based on, for example, parametric population parameter means or in the adaptive case, the MAP values of the parameters, the control is applied to all Npoint models from the population and some stochastic criterion is optimized – for example, the probability of hitting a target interval:**

M. Phillippe, M. Neely, Y. Bertrand, N. Bleysac, and S. Goutelle, A Nonparametric Method to Optimize Initial Drug Dosing and Attainment of a Target Exposure Interval: Concepts and Application Concepts and Application to Busulfan in Pediatrics, Clinical Pharmacokinetics, August 2016 (See the LAPK Website for further details on control theory applications with NP models)

- **Similar idea can be adapted to diagnostics – replace MAP-based diagnostics with their empirical Bayesian distribution analogs – M.Lavielle and B. Ribba, Enhanced Method for Diagnosing Pharmacometric Models: Random Sampling from Conditional Distributions, Pharm. Res. ,August 2016**

Historical objections to NP methods

- Speed – certainly NPML and NPEM were slower than contemporary parametric algorithms. This is emphatically no longer true – if special purpose optimization algorithms are used and/or the NP computation is piggybacked on top of a parametric optimization, NP phase is often much faster than the underlying parametric phase.
- Requirement to know residual error function in advance

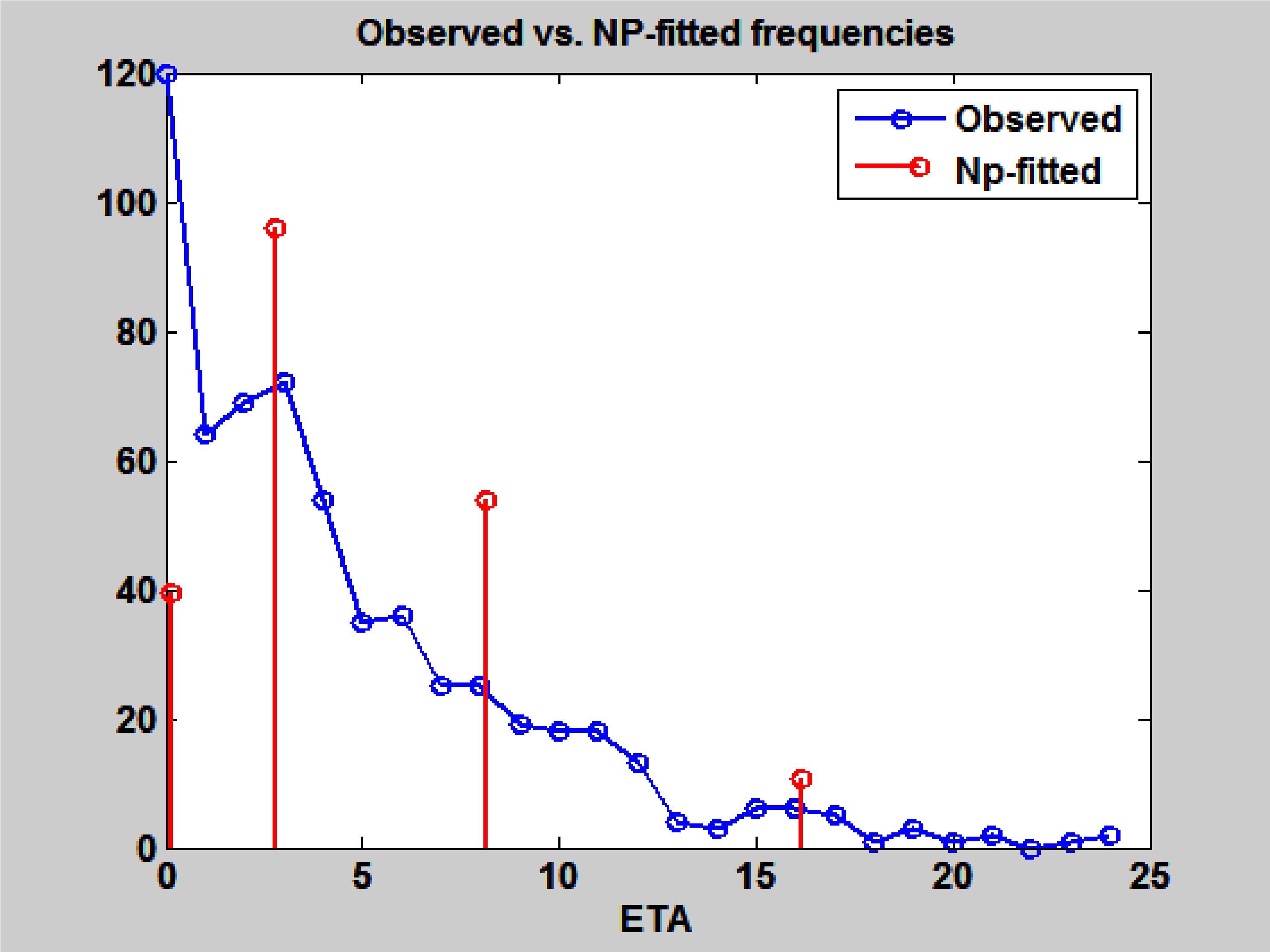
  Initial parametric run can be done to do this as in NOPD, Phoenix NLME- Residual error parameters (SIGMAs) can be estimated within NP (although only Pmetrics NPAG does this right now) by optimizing residual error function parameters on grid or simple profiling/
- Lack of standard error/uncertainty distribution capability –

  Standard errors and uncertainties of any functional of the population distribution (as well as individual estimates) can be obtained by bootstrap, which is much faster in the NP case than the parametric case.
- Lack of covariate optimization – can be handled either by mporting from iniital parametric run or by semi-parametric optimization .
- Discrete form of results

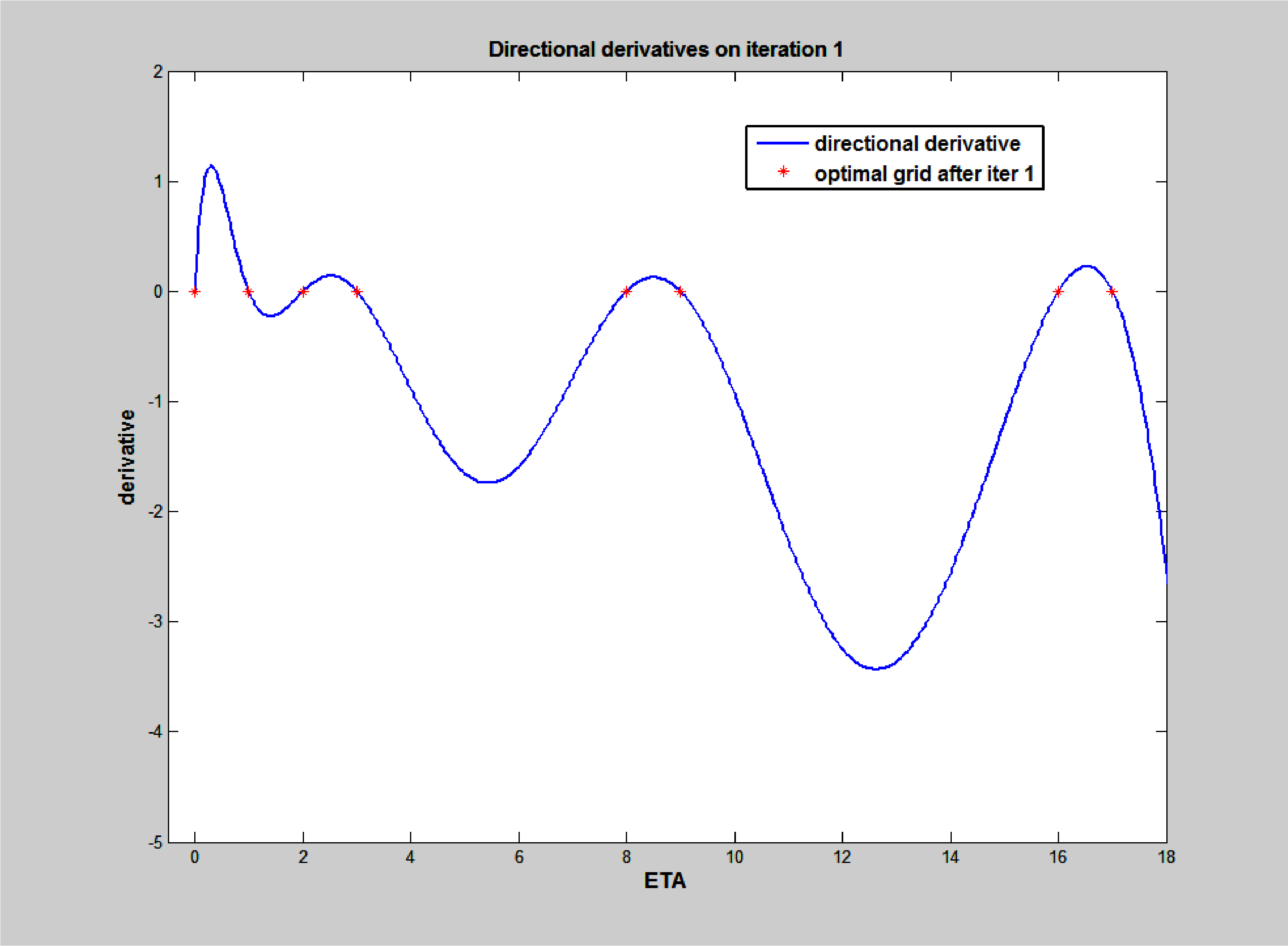
Advantages of Modern NP methods

- Speed – certainly NPML and NPEM were slower than contemporary parametric algorithms. This is emphatically no longer true – if special purpose optimization algorithms are used and/or the NP computation is piggybacked on top of a parametric optimization, NP is usually faster than the parametric 'accurate likelihood' methods.
- Likelihood evaluation – no approximations – for any given grid and probability vector, the likelihood is exact up to accuracy of underlying model evaluation at a single point. At grid optimization, likelihood always improves - much easier to evaluate convergence than accurate likelihood parametric methods.
- Reliability – easily the most reliable optimization among all commonly used methods – no numerical derivatives, resistant to local minima and saddle point termination
- Consistency – bias goes to zero as number of subjects increases, even for sparse data.
- Avoids model mis-specification at the random effect distribution level

Bohning PD example – count data



NPOD solution



Yong Wang comparison of speeds on simple Nsub=602 1-ETA case

Table 2. Computation of the NPMLE of the mixing distribution for the Thailand data set

Algorithm	*s*	$l(\hat{G}) - l(G_s)$	$sup_\theta\{d(\theta;G_s)\}$	*Time (s)*
VDM	70359	6.76×10^{-3}	1.00×10^{-2}	610.61
VEM	14064	6.14×10^{-3}	9.62×10^{-3}	434.06
EM	5337	1.85×10^{-7}	9.95×10^{-7}	5.58
ISDM	146	2.93×10^{-7}	5.24×10^{-7}	1.46
CN1O	68	5.17×10^{-8}	6.76×10^{-7}	1.35
CN1	56	6.34×10^{-8}	9.90×10^{-7}	0.84
CNM	20	1.18×10^{-9}	2.74×10^{-7}	0.27

Future work

- **Implement Non-negative least squares probability optimizer**
- **Implement NPOD or at least an improved pattern search method for grid expansion**
- **Implement fast bootstrap with NNLS**
- **Implement semi-parametric version**
- **Implement NP estimation within a PBPK framework**
- **Implement NP specific diagnostics**

Conclusions

- **NP estimation methods have entered the mainstream of POP PK/PD estimation methods – fast, accurate versions now widely available (NONMEM, Phoenix NLME, LAPK Pmetrics)**
- **Most commonly run after an initial parametric method, but can be run standalone**
- **Modern NP methods are very fast and reliable (often the fastest and most reliable of all current methods, with no likelihood approximations, no normality assumptions, no numerical derivatives, and very fast special purpose optimization algorithms)**
- **Grids based on just parametric EBE's are inferior to extended or iterated grids, particularly for sparse data, high-shrinkage cases**
- **Bootstrapping is particularly well suited to NP methods – re-solving a replicate just requires rerunning the probability computation with no new model evaluations**
- **NP methods naturally lead to 'multiple model' dose optimization and diagnostics – e.g. LAPK BestDose software.**
- **Need user feedback!**