An Introduction to Bayesian Estimation in NONMEM

Tim Waterhouse
February 22, 2023
Acknowledgements

- Curtis K. Johnston
- Matthew Wiens
- John Mondick
- Jonathan French
- Bill Gillespie

- “Tutorial: Bayesian Estimation in NONMEM” (manuscript submitted for publication)
- Introduction to Bayesian pharmacometric data analysis with NONMEM
  ○ ACoP 2019 workshop by Bill Gillespie and Curtis Johnston
Key messages

- Bayesian estimation can and should be used for pharmacometric models
- NONMEM speaks Bayes
- Some thought is required
  - Selecting priors
  - Processing output
Application of Bayesian approaches in drug development: starting a virtuous cycle

Stephen J. Ruberg, Francois Beckers, Rob Hemmings, Peter Honig, Telba Irony, Lisa LaVange, Grazyna Lieberman, James Mayne & Richard Moscicki

“... despite advances in Bayesian methodology, the availability of the necessary computational power and growing amounts of relevant existing data that could be used, Bayesian methods remain underused in the clinical development and regulatory review of new therapies.”
My prior belief in PMx use of Bayes

$p \sim \text{Beta}(1.5, 10)$

Overview

● Why Bayes?
● What is Bayesian analysis?
● Bayesian estimation in NONMEM
● Bayesian diagnostics
  ○ MCMC
  ○ Model
Why Bayes?
Why Bayes?

- Incorporate prior information
- Complexity in terms of random effects and hierarchies
- Analysis of data from heterogeneous sources
- Full posterior gives the best estimate of uncertainty
- Probabilistic inference for decision making
What is Bayesian analysis?

https://twitter.com/beth_fossen/status/1227244290763563008
Bayes’ Rule

\[
P(\theta, \Omega, \Sigma | y) = \frac{P(y | \theta, \Omega, \Sigma) P(\theta, \Omega, \Sigma)}{P(y)}
\]

Posterior distribution

Likelihood

Prior distribution

Data

(Purportedly) The Reverend Thomas Bayes
Bayes’ Rule

\[ P(\theta, \Omega, \Sigma | y) = \frac{P(y|\theta, \Omega, \Sigma)P(\theta, \Omega, \Sigma)}{P(y)} \]

- Posterior distribution
- Likelihood
- Prior distribution
- Data
- Prior
- Posterior
- Likelihood
Bayesian Modeling Computation

- Typically, no closed form posterior distribution
- Markov chain Monte Carlo (MCMC) used to sample from posterior
  - Metropolis-Hastings (MH)
  - Gibbs sampling
  - Hamiltonian Monte Carlo (HMC)
Prior distributions

- Represents prior knowledge or belief about model parameters
- Degrees of prior informativeness:
  - Informative
  - Weakly informative
  - Uninformative (e.g., uniform over positive real numbers)
- Explore with prior predictive simulation
Bayesian estimation in NONMEM
Control stream

● MU reference when possible
  ○ Allow Gibbs sampling (vs MH) for \texttt{METHOD=BAYES}
  ○ Analytic derivatives for \texttt{METHOD=NUTS}

● Prefer unbounded THETAs
  ○ Log or logit transform where possible

● Specify as many priors as possible
Control stream: Priors for THETAs

- Normal distribution
  - Mean $\text{THETAP}$
  - Variance $\text{THETAPV}$
    - Shorthand:
      $\text{THETAPV BLOCK(5) FIXED VALUES(10,0)}$

- t-distribution (METHOD=NUTS)
  - Set degrees of freedom in $\$EST \ TTDF$ or $\$TTDF$
Control stream: Priors for OMEGAs

● Inverse Wishart distribution

The probability density function of the inverse Wishart distribution is given by:

$$f_X(X; \Psi, \nu) = \frac{|\Psi|^{\nu/2} |X|^{-(\nu+p+1)/2} e^{-\frac{1}{2} tr(\Psi X^{-1})}}{2^{p/2} \Gamma_p \left(\frac{\nu}{2}\right)}$$

where $X$ and $\Psi$ are $p \times p$ positive definite matrices, $\cdot$ is the determinant, and $\Gamma_p(\cdot)$ is the multivariate gamma function.

Control stream: Priors for OMEGAs

- Inverse Wishart distribution
  - Mode $\Omega_{\text{MAP}}$
  - Degrees of freedom $\Omega_{\text{MAPD}}$
- Additional options for METHOD=NUTS:
  - Lognormal or half-t-distribution for SDs ($\text{EST OVARF}$)
  - Lewandowski-Kurowicka-Joe (LKJ) distribution for correlation matrix ($\text{EST OLKJDF}$)

where $X$ and $\Psi$ are $p \times p$ positive definite matrices, $\cdot$ is the determinant, and $\Gamma_p(\nu)$ is the multivariate gamma function.

Inverse Wishart OMEGA prior guidance

\[ df_i = CV(\Omega)^{-2} + n + 3 \]
\[ df = \min(df_i) \]
\[ \Omega_{\text{prior}} = \frac{df - n - 1}{df} E(\Omega) \]

where
- \( df_i = \) degrees of freedom for \( i^{th} \) OMEGA diagonal
- \( E(\Omega) = \) expected value of OMEGA diagonal
- \( CV(\Omega) = \) desired coefficient of variation for OMEGA diagonal
- \( n = \) number of diagonal elements in the OMEGA block

to set \( \Omega_{\text{prior}} (\text{OMEGAP}) \) and \( df (\text{OMEGAPD}) \)
Control stream: Priors for SIGMAs

- Inverse Wishart distribution
  - Mode $\text{SIGMAP}$
  - Degrees of freedom $\text{SIGMAPD}$

- Options for `METHOD=NUTS`:
  - Lognormal or half-t-distribution for SDs ($\text{EST SVARF}$)
  - Lewandowski-Kurowicka-Joe (LKJ) distribution for correlation matrix ($\text{EST SLKJDF}$)
Estimation options: initial estimates

- Multiple (e.g., 4) chains using `METHOD=CHAIN`.
  - Generate 4 sets of initial estimates with `METHOD=CHAIN NSAMPLE=4 FILE=1000.chn`.
  - Use `CTYPE` option to sample initial THETAs from
    - uniform (% above and below $\Theta$), or
    - bounds in $\Theta$ (not recommended!), or
    - normal distribution defined by $\Theta_P$ and $\Theta_{P\nu}$
  - OMEGA and SIGMA initial estimates from inverse Wishart distributions
    - Degrees of freedom from $DF$ and $DFS$
Estimation options: Sampling algorithm

- Metropolis-Hastings (MH) (*METHOD=BAYES*)
- Gibbs sampling (*METHOD=BAYES* with MU referencing)
- Hamiltonian Monte Carlo (HMC) (*METHOD=NUTS*)

MH is Meh
Gibbs is Good
HMC is How Maestros Compute
Estimation options: Individual posteriors

- **BAYES_PHI_STORE=1**
- Set of ETA samples for each draw from posterior
  - Provide individual-level summaries of uncertainty
  - Diagnostics (e.g., shrinkage, IPRED over full posterior)
Estimation options: Convergence testing

- **CTYPE=0**: no termination test (default, recommended)
- Tests based on changes in parameter estimates and/or objective function does not ensure convergence
Bayesian diagnostics
Bayesian diagnostics in NONMEM

- Diagnostics should consider full posterior (across all chains)
- NONMEM generates summaries (means, standard errors, shrinkages, etc.) within each chain
- Further post-processing is required to summarize and diagnose models across all chains
MCMC convergence diagnostics: graphical

- **Trace plots**
  - Check for stationary distribution with reasonable autocorrelation
    - ✅ fuzzy caterpillar
    - ❌ wiggly snakes

- **Density plots**
  - Common distribution between chains
MCMC convergence diagnostics: numerical

- **Rhat ($\hat{R}$)**
  - Measure of between-chain variance vs within-chain variance
  - Desire Rhat close to 1

- **Effective sample size (ESS)**
  - Measure of sampling efficiency
  - Bulk (location of distribution)
  - Tail (5th and 95th percentiles of distribution)
  - Desire ESS > ≈400

Bulk ESS = 188
Tail ESS = 354
Rhat = 1.01
MCMC convergence diagnostics: more graphical

- ESS vs draw
  - Will longer chains solve convergence issues?

- ESS vs quantile
  - Ensure convergence across all quantities of interest
Addressing convergence issues

- Constrain parameter space to plausible region using tighter priors
- For IIV parameters, sampling can also be improved by setting initial estimates for individual ETA values
  - E.g., single iteration of ITS first
- Reparameterize
  - Simplification: Consider identifiability
    - Non-Bayesian estimation can cover a multitude of sins
  - Non-centered parameterization
  - Truncated Emax
- Sampling algorithm: HMC (NUTS) > Gibbs > MH
Addressing convergence issues: Non-centered parameterization

- “Devil’s funnel” common in hierarchical (mixed effects) models
  - Sampler cannot explore sharp “neck”
- “Matt trick”:
  \[
  \begin{align*}
  x' &\sim N(0, 1) \\
  y' &\sim N(0, 1) \\
  x &= \exp(y/2) \cdot x' \\
  y &= 3 \cdot y'
  \end{align*}
  \]

- Set when `AUTO=2`

\[
\begin{align*}
  y &\sim N(0, 3) \\
  x &\sim N(0, \exp(y/2))
  \end{align*}
\]

https://mc-stan.org/docs/stan-users-guide/reparameterization.html
Model diagnostics
All typical PMx diagnostics apply

TUTORIAL

Model Evaluation of Continuous Data Pharmacometric Models: Metrics and Graphics

THT Nguyen1, M-S Moukassai2, N Holford3, N Al-Hunt4, T Freedman5, AC Hooker6, J John7, MO Karlsson8, DR Mould9, JJ Pérez Ruixo10, EL Plan11, R Savic11, JGC van Hasselt12, B Weber13, C Zhou14, E Comets1,18 and F Mentré1,18 for the Model Evaluation Group of the International Society of Pharmacometrics (ISoP) Best Practice Committee

This article represents the first in a series of tutorials on model evaluation in nonlinear mixed effect models (NLMEs), from the International Society of Pharmacometrics (ISoP) Model Evaluation Group. Numerous tools are available for evaluation of NLMEs, with a particular emphasis on visual assessment. This first basic tutorial focuses on presenting graphical evaluation tools of NLMEs for continuous data. It illustrates graphs for correct or misspecified models, discusses their pros and cons, and recalls the definition of metrics used.


- DV vs PRED/IPRED
- ETAs vs covariates
- VPCs
- NPDE vs time/covariates
- etc.
NONMEM output needs some tweaking

- NONMEM will output means of parameter estimates
  - Probably OK for THETAs, can introduce bias for variance terms
- $TABLE$ outputs
  - ETA values not derived across posterior distribution, but post hoc estimates using mean of THETAs/OMEGAs
    - May result in spurious correlations
Better to derive estimates/diagnostics using the full posterior: PREDs and IPREDs

- Simulate $S$ replicates:
  - $y_{ij}^{\text{sim,PRED}}$
    - include all variability, sample from posterior at each replicate
  - $y_{ij}^{\text{sim,IPRED}}$
    - include within-subject variability, posterior samples of population parameters and ETAs

- Calculate:
  - $\text{PRED}_{ij} = \frac{1}{S} \sum_{s=1}^{S} y_{ij}^{\text{sim,PRED}}$
  - $\text{IPRED}_{ij} = \frac{1}{S} \sum_{s=1}^{S} y_{ij}^{\text{sim,IPRED}}$
Better to derive estimates/diagnostics using the full posterior: Shrinkage

- Shrinkage = 1 – \( \frac{SD^K_{k=1}(\eta_k)}{\sqrt{\Omega}} \)
  - \( \eta_k \) is mean of ETA posterior samples for subject k
  - \( SD^K_{k=1} \) is standard deviation across K subjects
  - \( \Omega \) is mean of OMEGA estimates across posterior samples
Better to derive estimates/diagnostics using the full posterior: NPDE

- Can be calculated with npde R package
- Reuse output from PRED simulations:
  \[ y_{ijs}^{\text{sim}, \text{PRED}} \Rightarrow y_i^{\text{sim}(k)} \]

\[
Y_i^{\text{sim}(k)^*} = \text{var}(Y_i)^{-1/2}(Y_i^{\text{sim}(k)} - E(Y_i))
\]
\[
Y_i^* = \text{var}(Y_i)^{-1/2}(Y_i - E(Y_i))
\]
\[
pde_{ij} = F_i^{*}(y_{ij}^*) \approx \frac{1}{K} \sum_{k=1}^{K} \delta_{ijk}^*
\]

where \( \delta_{ijk}^* = 1 \) if \( y_{ij}^{\text{sim}(k)^*} < y_{ij}^* \) and 0 otherwise.

Key messages

- Bayesian estimation can and should be used for pharmacometric models
- NONMEM speaks Bayes
- Some thought is required
  - Selecting priors
  - Processing output
the end
Backup
Model selection criteria: what not to use

- Traditional objective function comparison not appropriate
- Alternatives: AIC, DIC, WAIC, cross-validation
  - **AIC**: not suitable for strong informative priors
  - **DIC**: unreliable for non-Gaussian posteriors
  - **WAIC**: not robust with weak priors or influential observations
  - **Cross-validation**: too computationally demanding
Model selection: Use PSIS-LOO

- **PSIS**: Pareto smoothed importance sampling
- **LOO**: leave-one-out cross-validation

\[
elpd_{\text{psis-loo}} = \sum_{i=1}^{n} \log \left( \frac{\sum_{s=1}^{S} w_i^s p(y_i|\theta^s)}{\sum_{s=1}^{S} w_i^s} \right)\]

- Available using **loo** R package
- \(p(y_i|\theta^s)\) is likelihood for a *subject* or *observation* at a given posterior sample
- Requires post hoc calculation with posterior ETAs

Prior impact assessment

- Sensitivity analysis: re-estimate with alternative priors
- Based on intended use of posterior inferences
- Alternatives should include changes to both variance and location of priors
- Goal is not to show limited impact, but to
  - Provide insight into any impact
  - Support use of prior to characterize external data
Example: Informative prior for $AUC_{50}$

![Graph showing density curves for different models](image)

- **Density**
- **PD Endpoint: Change from baseline**

**Model**
- Final Model
- Fixed (Variance to 0)
- 10x Variance
- 50x Variance
- 100x Variance
- Flat Prior
Example: Informative prior for $AUC_{50}$

Model
- Final Model
- Fixed (Variance to 0)
- 10x Variance
- 50x Variance
- 100x Variance
- Flat Prior

Model
- Final Model
- Extreme Large Mean
- 50% increased Mean
- 50% decrease Mean
- Extreme Small Mean