# FREQUENTIST BEHAVIOR OF FORMAL BAYESIAN INFERENCE

Donald A. Pierce
Oregon State Univ (Emeritus), RERF Hiroshima (Retired),
Oregon Health Sciences Univ (Adjunct)

Ruggero Bellio
Univ of Udine

For Perugia talk 7 July 2008
Slides are at ftp://www.science.oregonstate.edu/~piercedo/

# INTRODUCTION

We are interested in relations between frequentist inference and Bayesian inference using diffuse priors --- but differently than in terms of “matching priors”

Find that posterior densities can provide approximations to certain pseudo-likelihoods, providing for quite accurate frequentist inference from the Bayesian calculations

Particularly interested in Bayesian MCMC methods, useful when evaluating the likelihood function involves intractable integrals; e.g. GLMM, covariate errors, imputation of missing data, spatial (and image) analysis, and so forth

Many of those using this seem really to be seeking “frequentist” inference, only using the MCMC to resolve difficulties in evaluating the likelihood

For very large *n,* frequentist and Bayesian inference coincide since $(\hat{\theta}-\theta)/SE$ has the same Gaussian limiting distribution from either perspective

Not so “comforting” if either of these distributions is clearly not Gaussian

Modern frequentist likelihood asymptotics points to deeper connections between the two modes of inference

This pertains largely to the treatment of *nuisance parameters*

Usually, in frequentist methods these are “maximized out” of the likelihood function, but in Bayesian methods they are “integrated out”

But modern (frequentist) likelihood asymptotics suggests also the “integrating out” of nuisance parameters

Inference about $\psi(\theta)$ where $\theta$ is multi-dimensional

Reparametrize as $(\psi, \lambda)$, realizing that the choice of representation of $\lambda$ is largely arbitrary

Profile likelihood (*PL*) is fundamental and does not depend on that arbitrary matter

$$L_P(\psi) = \max_\lambda L(\psi, \lambda; y) = \max_{\psi(\theta)=\psi} L\big(\theta; y\big)$$

Modern likelihood asymptotics points to two types of modification of *PL,* dealing with effects of fitting nuisance parameters

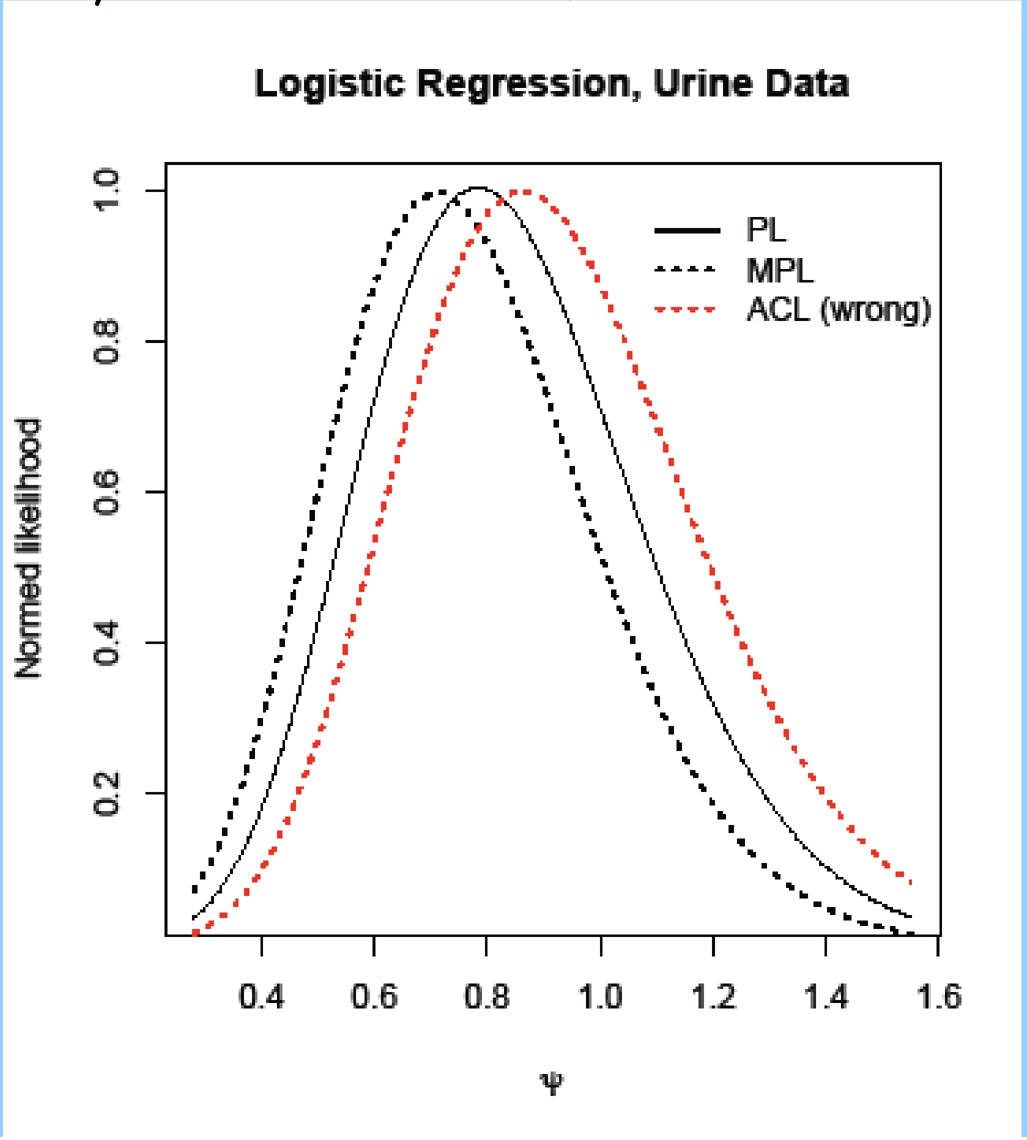
Approximate conditional likelihood (*ACL*) in terms of “orthogonal” nuisance parameter $\lambda^\perp$ (not unique)

$$L_{AC}(\psi) = L_P(\psi)\ |\,\mathrm{var}(\hat{\lambda}^\perp; \psi\ \mathrm{known})\,|^{1/2}$$ (using asy var, obsvd info)

Modified profile likelihood (*MPL*) which does not depend on representation of nuisance parameters

$$L_{MP}(\psi) = L_{AC}(\psi)\ |\,\partial\hat{\lambda}^\perp / \partial(\hat{\lambda}^{|\psi})\,|$$ ( $\hat{\lambda}^{|\psi}$ constrained MLE)

Logistic regr: 77 binary obsvns with 6 covars. The "ACL" here uses the formula for that but with the non-orthogonal canonical parameter. When using the orthogonal mean parameter, the ACL is the same as the MPL



Orthogonal parameters: Main point is that the MLE of $\lambda^{\perp}$ when $\psi$ is known does not depend strongly on $\psi$

Will be true if the expected information matrix for $(\psi, \lambda)$ is block diagonal

Leads to a PDE with solution giving $\lambda^{\perp}$, always *exists* when $\psi$ is scalar --- $\lambda^{\perp}$ depends on both $\lambda$ and $\psi$ (unless …)

The above Jacobian $|\partial\hat{\lambda}^{\perp}/\partial(\hat{\lambda}^{|\psi})|$ is a quantity whose precise definition is complicated, but . . .

Can approximate it adequately in terms of the covariance of the score vector at two different parameter values --- simpler approximation when the data can be represented in stochastically independent subsets

Orthogonal parameters are important to what follows, with interesting Bayesian and MCMC implications

*ACL* based on common approach of conditioning to eliminate nuisance parameters

It depends on the non-unique choice of the orthogonal parameter

*MPL* invariant to that choice, approximating conditional likelihood when that is appropriate and “marginal” likelihood when that is appropriate

* It appears that *MPL* is generally approximating an idealized “integrated” likelihood *IL*

$$L_I^{ideal}(\psi) = \int L(\psi, \lambda; y)\, w(\lambda) d\lambda$$

where the weight function $w(\lambda)$ is, in principle, a suitably invariant distribution on $\psi^{-1}(\psi)$ --- difficult to deal with explicitly

Severini (2007) developed a good approximation to that aim, but essentially with a complicated weight function $w(\lambda)$ depending on the data

Bayesian calculations: If $\lambda, \psi$ are *a priori* independent then

$$\pi(\psi \mid y) \propto \int L(\psi, \lambda; y)\pi(\psi)d\pi(\lambda) = \pi(\psi)\int L(\psi, \lambda; y)d\pi(\lambda)$$

so

$$\pi(\psi \mid y) \propto \pi(\psi)\, L_I^{\lambda,\pi}(\psi; y)$$

where the special *IL* depends on the representation of the nuisance parameter and the prior for it (related matters)

Note that the prior $\pi(\psi)$ has a very simple effect, not very important for our considerations, since we are mainly concerned with some kind of pseudo-*likelihood* for $\psi$

An important Laplace approximation (Sweeting 1987) is

$$L_I^{\lambda,\pi}(\psi; y) \doteq L_P(\psi)\ |\mathrm{var}(\hat{\lambda}; \psi\ \mathrm{known})|^{1/2}\ \pi(\hat{\lambda}^{|\psi} \mid \psi)$$

which is formula for *ACL* but without imposing orthogonality

An *IL* using an orthogonal parameter may approximate well a version of *ACL*, and perhaps even the *MPL*

Often fairly easy with MCMC to obtain samples from

$$\pi(\psi, \lambda \mid y) \propto \pi(\psi)\pi(\lambda)\, p(y \mid \psi, \lambda)$$

As noted above, if $\pi(\psi)$ is uniform then the marginal density of $\psi$ from this is an integrated likelihood $L_I^{\lambda,\pi}(\psi; y)$ and this will depend little on the choice of a diffuse $\pi(\lambda)$

Recall that for diffuse $\pi(\lambda)$

$$L_I^{\lambda,\pi}(\psi; y) \doteq L_P(\psi)\ \mid \mathrm{var}(\hat{\lambda}; \psi\,\mathrm{known}) \mid^{1/2}$$

but this is not *ACL* unless $\lambda$ is orthogonal

Seems impractical to ask that the MCMC be done, generally, in an orthogonal parameterization

It seems not so difficult approximate the final factor from analysis of the MCMC samples, thus to “remove” it and obtain *PL*

We suspect, but less confidently, that one can generally approximate from the samples a desired adjustment factor

$$\mid \mathrm{var}(\hat{\lambda}^{\perp}; \psi\,\mathrm{known}) \mid^{1/2}$$

In the example given above, the MCMC posterior density is essentially the same as the “wrong ACL” when, as usual, $\lambda$ is represented in terms of canonical parameters

If the MCMC were done taking $\lambda$ as the orthogonal mean parameter, the posterior density would be essentially the same as the MPL.

So one can obtain useful likelihood quantities from the MCMC, but it is crucial to use something close to the orthogonal representation of nuisance parameters

How to do this can be worked out for some important classes of problems, e.g. GLMMs, but it is difficult to do in general, so we need a way to “fix up” the damage from using non-orthogonal parameters.

Some comments:

There is much written about choice of $\pi(\psi)$ but for our intentions this is not really an issue

If not taken as uniform, then can divide by it at the end. Doing this may affect the convergence of the MCMC

It is not really necessary to “sample” $\psi$ at all --- can take a grid of values for it and only sample $\lambda$ . In WinBUGS, say, make many separate runs with $\psi$ fixed

As for orthogonal parameters, it will not work simply to transform the samples after the simulation --- there are complications regarding the Jacobian that possibly could be dealt with

Even when a form of orthogonal parameter is known in simple form, it can be difficult to use that in the MCMC specifications

The quantities $\log|\,\mathrm{var}(\hat{\lambda};\psi\,\mathrm{known})\,|^{1/2}$ are very smooth functions, often essentially linear in $\psi$

Example: Logistic regression with random intercept for clusters.

$$\log\text{it}(p_{ij}) = z_{ij}\beta + \sigma u_i$$

When taking $\psi = \sigma$ , it turns out that a nearly orthogonal parametrization replaces $\beta$ by $\beta^{\perp} = \beta / \sqrt{1+0.3\sigma^2}$

This derives from considering the usual "attenuation" in marginal estimation with mixed models

Things work fine, i.e. the WinBUGS posterior density for $\sigma$ is essentially the *MPL* as desired: the problem is that rather special considerations were necessary to obtain $\beta^{\perp}$
(When $\psi = \beta_j$ the original lack of orthogonality is minor)

Essentially the same approach works for any GLMM with random intercepts for clusters, and for identity, log or logit link (as above).

But, can we fix things up regarding non-orthogonal parameters in a more general manner?

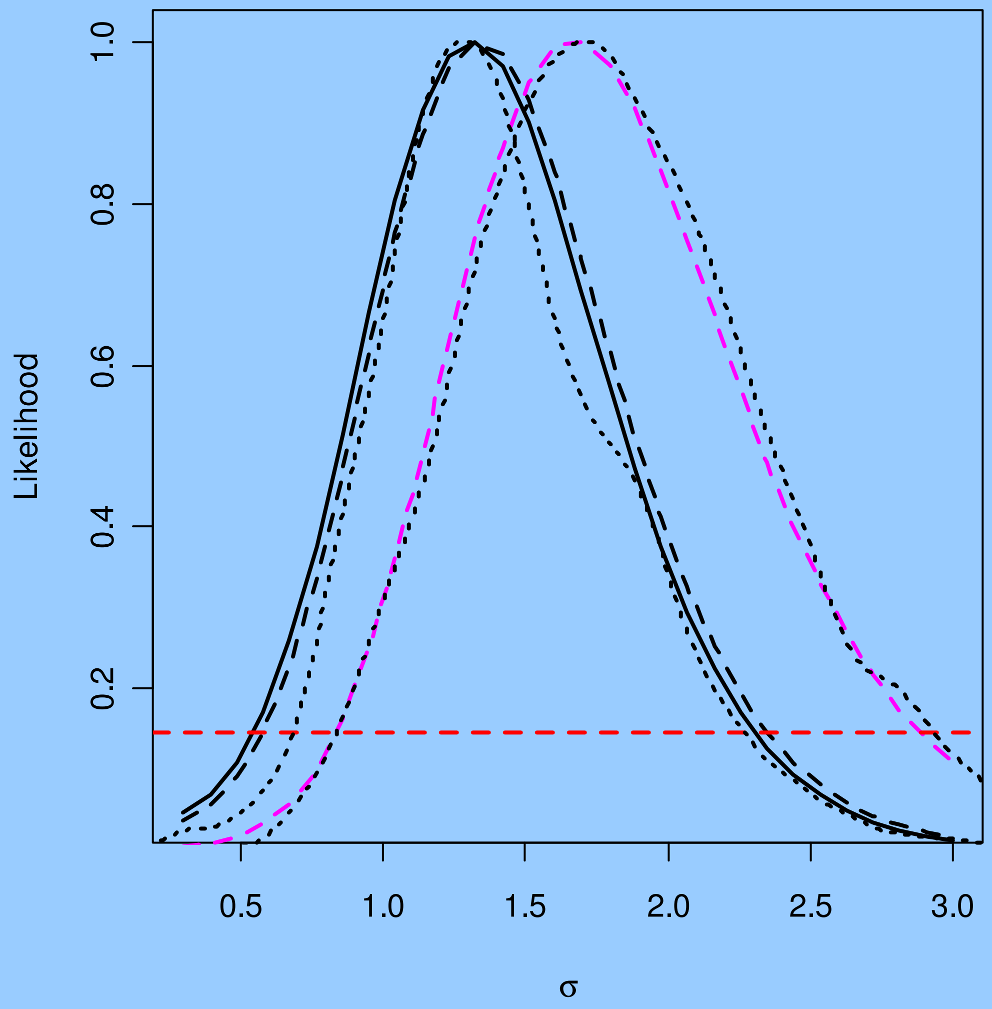
“Bacteria” data --- binary data on 50 individuals, repeated observations at 0,2,4,6,11 weeks (some missing) for 220 total binary readings.

Model: 3 parameters for treatments, 4 parameters for time variation

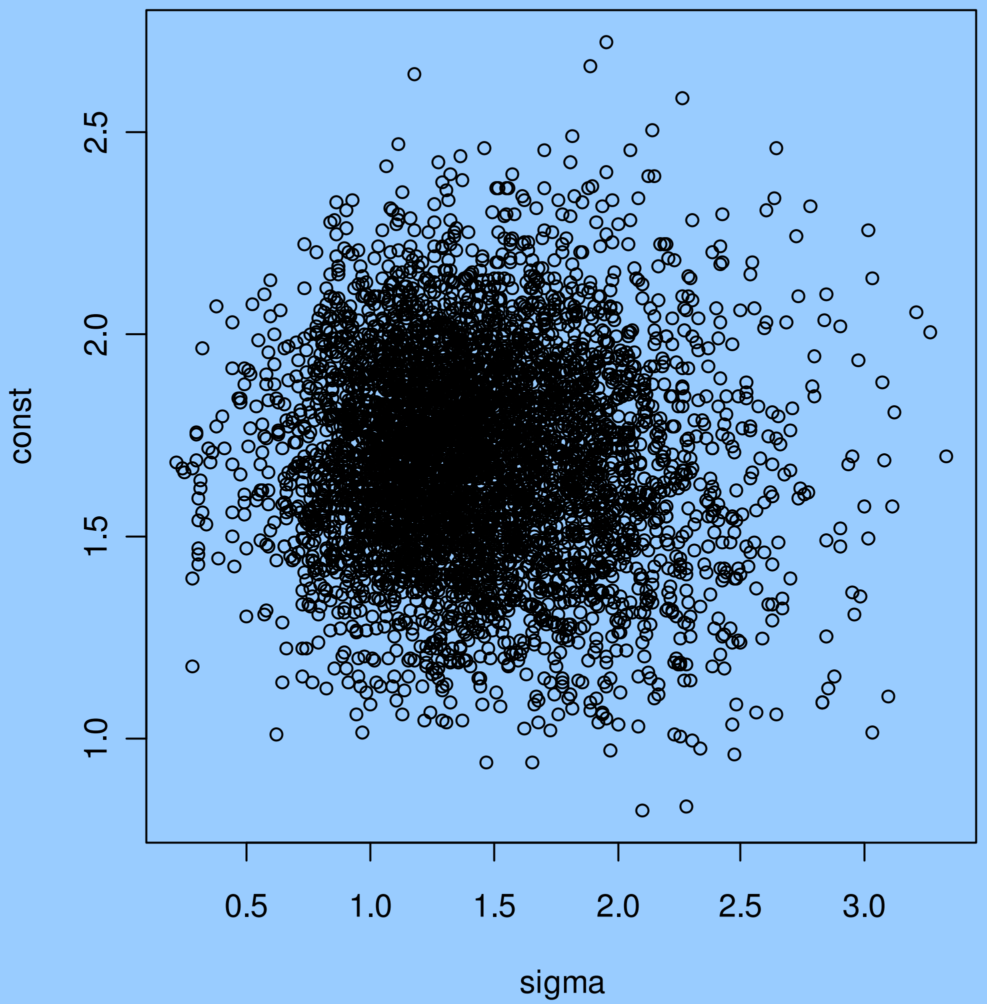
Logistic mixed model with $N(0,\sigma^2)$ term on logit scale for each individual, to allow for within-individual correlation

WinBUGS analysis, and HOA likelihood analysis using R-program prepared by Bellio in work with Brazzale

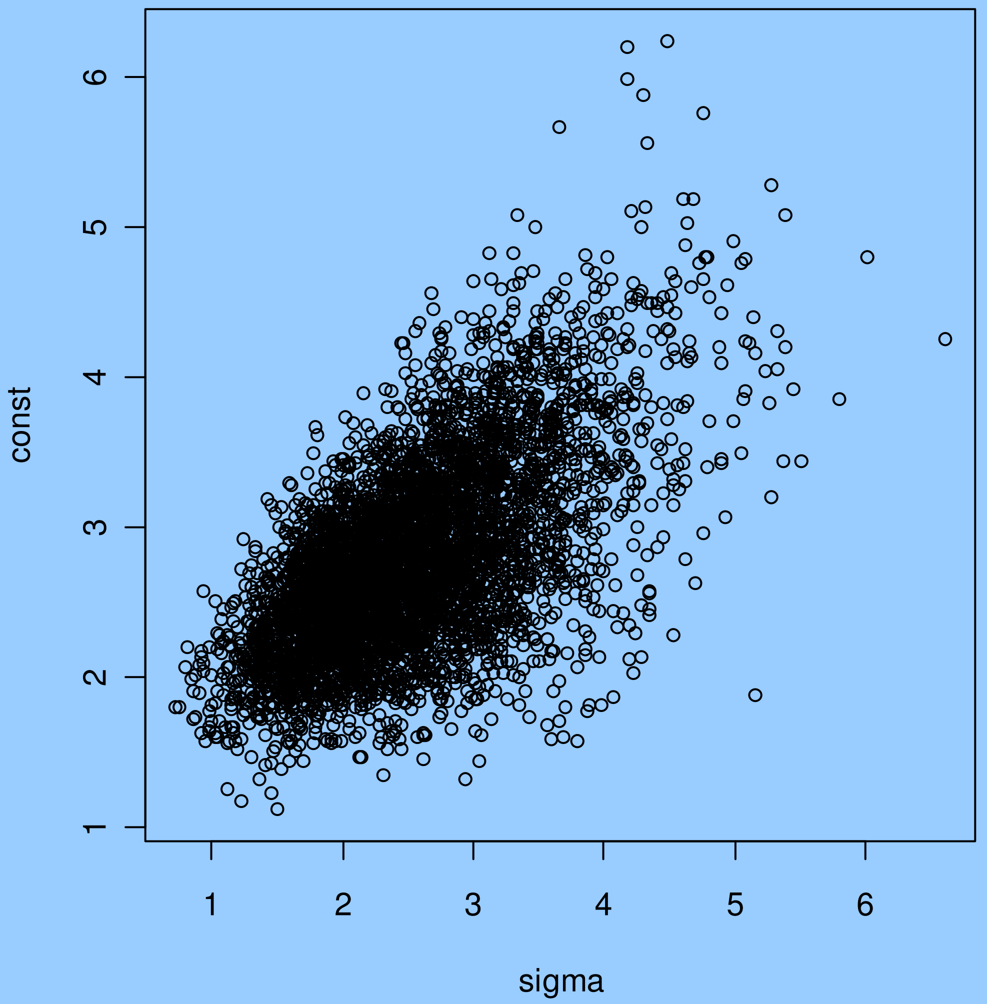
Consider here only inference about sigma



3 left curves: PL, MPL, WinBUGS posterior orthog param
2 right curves: APL formula (Laplace), WinBUGS orig param



Posterior samples (sigma, const) in orthog params: Orthog seen as constant regression. Variation of Var(const|sigma) would provide for adjustment PL to APL, but none seen



Original params: Lack of orthog seen as non-constant regression. Var(const|sigma) varies too much with sigma, including unwanted variation due to E(const|sigma)

Recall that

$$L_I^{\lambda,\pi}(\psi; y) \doteq L_P(\psi)\ |\operatorname{var}(\hat{\lambda}; \psi\, \text{known})|^{1/2}$$

and consider approximating the final factor from the MCMC samples, so we can (first) remove it to obtain an approximation to *PL*

From the MCMC samples we can compute (estimate arbitrarily well)

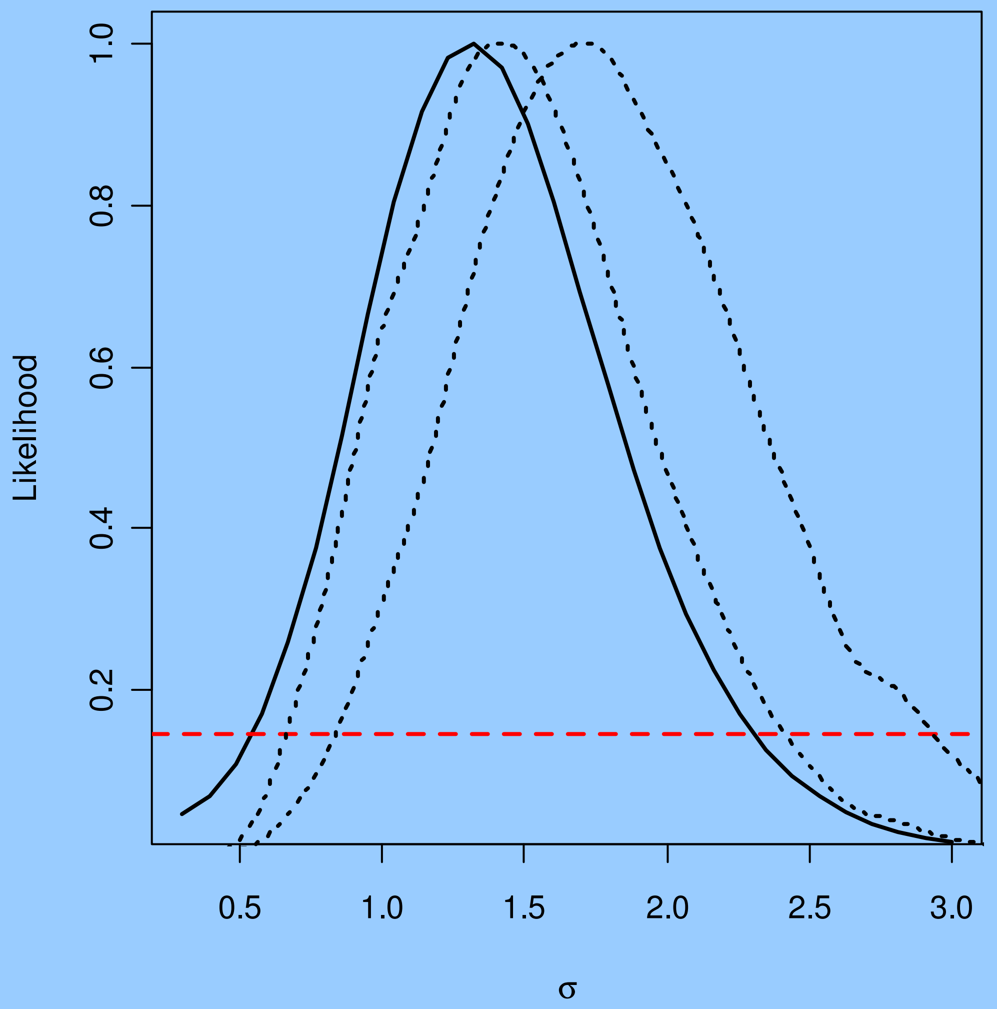
$$\operatorname{var}(\tilde{\lambda} \,|\, y\ ; \psi\ known) = \operatorname{var}(\tilde{\lambda} \,|\, \tilde{\psi} = \psi, y)$$

(recall that for our needs we could either fix $\psi$ at specified values, or sample it from $\pi(\psi)$ )

With diffuse prior we can expect that

$$\operatorname{var}(\hat{\lambda}; \psi\ known) \doteq \operatorname{var}(\tilde{\lambda} \,|\, y; \psi\ \ known)$$

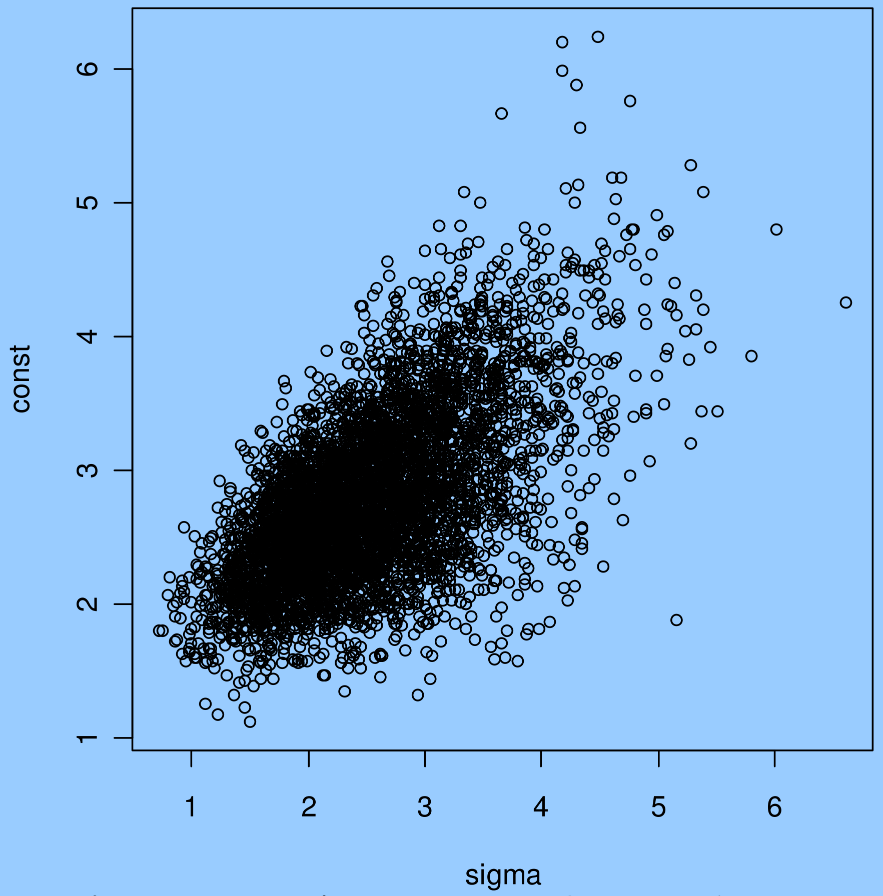
Making this substitution, estimating the conditional variance function from posterior samples, and "back-calculating" the *PL* works quite well in various examples we have tried



Profile (solid), posterior density (rightmost dashed), and posterior adjusted by estimated variance function (leftmost dashed) ---- recovering essentially the *PL* from the MCMC. Can actually do better than this with modification of details

Can we go further and approximate an *ACL* or the *MPL* ?

Need to estimate the function $\mathrm{var}(\tilde{\lambda}^{\perp} \mid \tilde{\psi} = \psi)$



One possibility is to “remove” from the variance function here that variation due to the unwanted dependence $E(\tilde{\lambda} \mid \tilde{\psi} = \psi)$

Another is to estimate more directly from the posterior samples key aspects of the transformation to orthogonal parameters

Possible resolution of these issues is currently underway

So where are we so far?

We can use MCMC to approximate the *PL* and likely an *ACL or* MPL

Distinctions between using the posterior as a density and using it as a likelihood are readily explored (often will not matter much)

Can obtain very good frequentist inferences without complications of choosing prior for interest parameter

As a practical matter, provides for frequentist inference when the likelihood is difficult to evaluate

As a theoretical matter, clarifies connections between the two modes of inference