Probability Matching Priors in LHC Physics: A Pragmatic Approach

Paul Baines (joint work with Xiao-Li Meng)

Department of Statistics
Harvard University

August 1, 2007

Context

For those of you who saw Jim Berger's first Wald Lecture yesterday...

This talk is related to the LHC Physics problem he discussed – the upper confidence limits part...

...and is the same setting as Paul Edlefsen's talk (first talk of this session)

Motivating Example

Consider the following common problem arising in LHC Physics:

$$
\begin{aligned}
n_i &\sim \mathrm{Pois}(\epsilon_i s + b_i) \\
y_i &\sim \mathrm{Pois}(t_i b_i) \\
z_i &\sim \mathrm{Pois}(u_i \epsilon_i)
\end{aligned}
$$

with $i = 1, \ldots, M$ indexing the decay channels.

- s: The Poisson rate of ‘source’ counts (common to all channels)
- b: The Poisson rate of ‘background’ counts per channel
- ϵ: The decay rate per channel.

Motivating Example

Consider the following common problem arising in LHC Physics:

$$\begin{aligned} n_i &\sim \text{Pois}(\epsilon_i s + b_i) \\ y_i &\sim \text{Pois}(t_i b_i) \\ z_i &\sim \text{Pois}(u_i \epsilon_i) \end{aligned}$$

with $i = 1, \ldots, M$ indexing the decay channels.

- s: The Poisson rate of 'source' counts (common to all channels)
- b: The Poisson rate of 'background' counts per channel
- ϵ: The decay rate per channel.

In this talk we focus on $M = 1$ & $M = 10$: the single & ten-channel cases.

The Problem

Goal:

Find a method for producing ***'reliable interval estimates'*** *for a univariate parameter of interest (i.e.* s*) in the presence of nuisance parameters (i.e.* b, ϵ*)*

Coverage Criterion

One-sided: $$\mathbb{P}_\theta\left(s < s^{(1-\alpha)} | s, b, \epsilon\right) = 1 - \alpha$$

where $\mathbb{P}$ is the Frequentist probability measure and $s^{(1-\alpha)}$ is the $(1-\alpha)^{\text{th}}$-percentile produced by a given method i.e. a data-dependent random variable.

Coverage Criterion

One-sided: $$\mathbb{P}_{\theta}\left(s < s^{(1-\alpha)} | s, b, \epsilon\right) = 1 - \alpha$$

where $\mathbb{P}$ is the Frequentist probability measure and $s^{(1-\alpha)}$ is the $(1-\alpha)^{\text{th}}$-percentile produced by a given method i.e. a data-dependent random variable.

Two-sided: $$\mathbb{P}\left(s \in S^{(1-\alpha)} | s, b, \epsilon\right) = 1 - \alpha$$

where $S^{(1-\alpha)}$ is the set of s values contained in the $(1-\alpha)^{\text{th}}$-percentile interval produced by a given method i.e. a data-dependent random interval. e.g. $(s^{(0.025)}, s^{(0.975)})$.

'Default' priors

Here we focus on Bayesian approaches. There are a plethora of 'default' or 'non-subjective' priors in the literature, including:

'Default' priors

Here we focus on Bayesian approaches. There are a plethora of 'default' or 'non-subjective' priors in the literature, including:

1. **Flat Priors:** (Laplace) Notorious problems
2. **Jeffrey's Prior:** Problems in multi-dimensions
3. **Probability Matching Priors:** More later...
4. **Reference Priors:** More later...
5. **Trade-Off Priors:** (Clarke & Wasserman, 1993)
6. **Haar Measures:** Based on invariance considerations
7. **MDIP Priors:** (Zellner, 1971)
8. **Indifference Prior:** (Novick & Hall, 1965)

'Default' priors

Here we focus on Bayesian approaches. There are a plethora of 'default' or 'non-subjective' priors in the literature, including:

1. **Flat Priors:** (Laplace) Notorious problems
2. **Jeffrey's Prior:** Problems in multi-dimensions
3. **Probability Matching Priors:** More later...
4. **Reference Priors:** More later...
5. **Trade-Off Priors:** (Clarke & Wasserman, 1993)
6. **Haar Measures:** Based on invariance considerations
7. **MDIP Priors:** (Zellner, 1971)
8. **Indifference Prior:** (Novick & Hall, 1965)

These are not distinct classes of priors, they frequently coincide.

An Aside. . .

The challenge posed by the Physicists is essentially to provide a baseline solution. The 'best' prior for this purpose may not be the preferred prior for actual data analysis though.

An Aside...

The challenge posed by the Physicists is essentially to provide a baseline solution. The 'best' prior for this purpose may not be the preferred prior for actual data analysis though.

It may be of interest to combine subjective priors for nuisance parameters with a 'non-subjective' prior for the interest parameter. See Demortier (2005) for details of this within the reference prior framework.

Since the Physicists primary interest is in coverage we focus on...

Probability Matching Priors

Probability Matching Priors (PMP) are a bridge between Bayesian and Frequentist methodologies (with some qualifications).

Probability Matching Priors

Probability Matching Priors (PMP) are a bridge between Bayesian and Frequentist methodologies (with some qualifications).

- Provide posterior intervals with Frequentist validity
- Can be used as a formal rule for selecting the prior distribution
- Can be used as a constructive tool for Frequentist inference (e.g. Levine & Casella, 2003)

Formal Definition

Definition

(Exact) Probability Matching Prior: *Let* $\{f(\cdot|\theta) : \theta \in \Theta\}$ *be a parametric family where* $\theta = (\psi, \phi) \in \mathbb{R}^p$*. Let* $\psi \in \mathbb{R}$ *be the parameter of interest, with* $\phi \in \mathbb{R}^{p-1}$ *considered to be a* $(p-1)$*-dimensional nuisance parameter. Let* $\psi^{(1-\alpha)}(\pi, \mathbf{Y})$ *denote the* $100(1-\alpha)^{th}$ *(marginal) posterior percentile for* ψ *with observed data* $\mathbf{Y}$*, and under the prior* π*. A prior distribution* $\pi(\theta)$ *is said to be (exact) probability matching for* ψ *if:*

$$\mathbb{P}_\theta\left(\psi \leq \psi^{(1-\alpha)}(\pi, \mathbf{Y})\right) = 1 - \alpha \qquad (1)$$

Formal Definition

Definition

r^{th} **Order Probability Matching Prior**: *Let* $\{f(\cdot|\theta) : \theta \in \Theta\}$ *be a parametric family where* $\theta = (\psi, \phi) \in \mathbb{R}^p$. *Let* $\psi \in \mathbb{R}$ *be the parameter of interest, with* $\phi \in \mathbb{R}^{p-1}$ *considered to be a* $(p-1)$*-dimensional nuisance parameter. Let* $\psi^{(1-\alpha)}(\pi, \mathbf{Y})$ *denote the* $100(1-\alpha)^{th}$ *(marginal) posterior percentile for* ψ *with observed data* $\mathbf{Y}$*, and under the prior* π*. A prior distribution* $\pi(\theta)$ *is said to be* r^{th} *order probability matching for* ψ *if:*

$$\mathbb{P}_\theta\left(\psi \leq \psi^{(1-\alpha)}(\pi, \mathbf{Y})\right) = 1 - \alpha + o(n^{-r/2}) \tag{2}$$

The Reward

Some background:

The Reward

Some background:

- There are a plethora of methods in the Physics literature: some standard (e.g. profile likelihood), some *ad hoc* (e.g. hybrid Frequentist-Bayes)

The Reward

Some background:

- There are a plethora of methods in the Physics literature: some standard (e.g. profile likelihood), some *ad hoc* (e.g. hybrid Frequentist-Bayes)
- This led to the 'Banff Challenge'! (see Paul Edlefsen's talk)

The Reward

Some background:

- There are a plethora of methods in the Physics literature: some standard (e.g. profile likelihood), some *ad hoc* (e.g. hybrid Frequentist-Bayes)
- This led to the 'Banff Challenge'! (see Paul Edlefsen's talk)
- PMPs offer an 'optimal' solution (when they exist, and up to the desired order of approximation. . .)

The Reward

Some background:

- There are a plethora of methods in the Physics literature: some standard (e.g. profile likelihood), some *ad hoc* (e.g. hybrid Frequentist-Bayes)
- This led to the 'Banff Challenge'! (see Paul Edlefsen's talk)
- PMPs offer an 'optimal' solution (when they exist, and up to the desired order of approximation. . .)
- There often exists a **class** of PMP's $\Rightarrow$ select on other criteria

The Reward

Some background:

- There are a plethora of methods in the Physics literature: some standard (e.g. profile likelihood), some *ad hoc* (e.g. hybrid Frequentist-Bayes)
- This led to the 'Banff Challenge'! (see Paul Edlefsen's talk)
- PMPs offer an 'optimal' solution (when they exist, and up to the desired order of approximation...)
- There often exists a **class** of PMP's $\Rightarrow$ select on other criteria
- Accessible to both Frequentist and Bayesians Physicists!

Characterization Theorem I

Theorem

(Peers, 1965) First Order Matching Prior Condition:

(A) *A prior* $\pi(\psi, \phi)$, $\psi \in \mathbb{R}$, $\phi \in \mathbb{R}^{p-1}$ *is first order probability matching for* ψ *if and only if it satisfies the PDE:*

$$\frac{\partial}{\partial \psi}\left\{\pi(\psi,\phi)\cdot\left(I^{\psi\psi}\right)^{1/2}\right\} + \sum_{j=1}^{p-1}\frac{\partial}{\partial \phi_j}\left\{\pi(\theta) I^{\phi_j\psi}\left(I^{\psi\psi}\right)^{-1/2}\right\} = 0 \qquad (3)$$

where I^{ij} *is the entry of the inverse Fisher Information matrix corresponding to the parameters* (i, j).

Characterization Theorem II

Theorem

(Mukerjee & Ghosh, 1997) Second Order Matching:

(B) *The prior $\pi(\cdot)$ is also second probability matching for ψ if and only if it satisfies the additional PDE:*

$$\sum_{j=0}^{p-1}\sum_{r=0}^{p-1}\left\{\frac{\partial}{\partial\phi_j}\frac{\partial}{\partial\phi_r}\left[\pi(\theta)\left(\frac{I^{\phi_j,\psi}I^{\phi_r,\psi}}{I^{\psi,\psi}}\right)\right]-\right.$$

$$\frac{1}{3}\sum_{u=0}^{p-1}\sum_{s=0}^{p-1}\frac{\partial}{\partial\phi_u}\frac{\partial}{\partial\phi_s}\left[\pi(\cdot)\left(\frac{I^{\phi_j,\psi}I^{\phi_r,\psi}}{I^{\psi,\psi}}\right)\mathbb{E}_\theta\left[\frac{\partial^3}{\partial\phi_j\partial\phi_r\partial\phi_s}\log f(Y_1;\psi,\phi)\right]\cdot\right.$$

$$\left.\left.\left\{3\left[I^{\phi_s\phi_u}-\left(\frac{I^{\phi_s,\psi}I^{\phi_u,\psi}}{I^{\psi,\psi}}\right)\right]+\left(\frac{I^{\phi_s,\psi}I^{\phi_u,\psi}}{I^{\psi,\psi}}\right)\right\}\right]\right\}=0$$

where ϕ_0 is defined to be ψ for notational convenience.

Challenges

- Potentially high-dimensional and non-linear PDE
- Analytic solutions rarely possible
- Standard software for solving PDE's (Mathematica, Maple) can rarely solve these equations (even numerically in many cases)
- Where solutions are possible, parts of the prior are often specified only up to an arbitrary function.
  (This can be dealt with though. . .)

Existing Approaches

To date, there appear to only have been two (direct) approaches to find a framework for implementing PMP's:

Existing Approaches

To date, there appear to only have been two (direct) approaches to find a framework for implementing PMP's:

(1) **Levine & Casella (2003):** Numerical solution of PDE's via method of characteristics, embedded in an MCMC scheme.

- Specific to $p = 2$ setting (i.e. univariate nuisance parameter)

Jump to orthogonality/reference priors...

Existing Approaches

To date, there appear to only have been two (direct) approaches to find a framework for implementing PMP's:

(1) **Levine & Casella (2003):** Numerical solution of PDE's via method of characteristics, embedded in an MCMC scheme.
- Specific to $p = 2$ setting (i.e. univariate nuisance parameter)

(2) **Sweeting (2005):** Seek *local probability matching priors*, using data-dependent approximations.
- More generally applicable, but requires a non-trivial condition on the parameterization

Jump to orthogonality/reference priors...

Existing Approaches

To date, there appear to only have been two (direct) approaches to find a framework for implementing PMP's:

(1) **Levine & Casella (2003):** Numerical solution of PDE's via method of characteristics, embedded in an MCMC scheme.
   - Specific to $p = 2$ setting (i.e. univariate nuisance parameter)

(2) **Sweeting (2005):** Seek *local probability matching priors*, using data-dependent approximations.
   - More generally applicable, but requires a non-trivial condition on the parameterization
   - Both are recent work (no applications of either method published to date)

Jump to orthogonality/reference priors...

Alternative Approaches

Recall that $I^{\psi,\phi}$ are the coefficients in the PMP PDE. What if a parameterization is 'almost orthogonal'?

Alternative Approaches

Recall that $I^{\psi,\phi}$ are the coefficients in the PMP PDE. What if a parameterization is 'almost orthogonal'?

If the structure of the prior remains largely determined by the first term (with $I_{\psi,\psi}$ coefficient) then, subject to a certain 'smoothness' of the PDE, we may expect the coverage properties of 'orthogonal' PMP's to be 'good'.

$$\text{i.e.} \qquad \pi(\psi,\phi) \propto \sqrt{I_{\psi,\psi}(\psi,\phi)}$$

Orthogonality Index

The concept of being 'almost orthogonal' can be made rigorous. We propose the following criteria. . .

Relative Information (RI)

Definition

Relative Information: *Consider a parameterization $\theta = (\psi, \phi)$ with ψ interest, ϕ nuisance. Denote the elements of the partitioned Fisher Information matrix by I_{ij}, $i, j = \psi, \phi$. Define the relative information (RI) for ψ in the $\theta-$parameterization to be:*

$$RI(\theta) := \frac{I_{\psi,\psi}(\theta) - I_{\psi,\phi}(\theta)\left(I_{\phi,\phi}(\theta)\right)^{-1} I_{\phi,\psi}(\theta)}{I_{\psi,\psi}(\theta)} \qquad (4)$$

Orthogonality Index (OI)

Definition

Orthogonality index: *Consider a parametric family* $\{f_\theta(\cdot)\}$, $\theta \in \Theta$, *with* $\theta = (\psi, \phi)$. *The orthogonality index of the parameterization* θ *with respect to the measure* π *is defined to be, for* $\dim(\psi) = 1$*:*

$$OI_{f_\theta}(\pi) := \mathbb{E}_\pi\left[RI(\theta)\right]$$

$$OI_{f_\theta}(\pi) := \int_\Theta \frac{I_{\psi,\psi}(\theta) - I_{\psi,\phi}(\theta)\left(I_{\phi,\phi}(\theta)\right)^{-1} I_{\phi,\psi}(\theta)}{I_{\psi,\psi}(\theta)} \pi(\theta) d\theta$$

Multivariate *OI* Definition

The extension to the $\dim(\psi) = p$ case is taken to be:

$$OI_{f_\theta}(\pi) := \int_\Theta \left(\mathbb{I}_p - I_{\psi,\psi}^{-1/2}(\theta) I_{\psi,\phi}(\theta) \left(I_{\phi,\phi}(\theta) \right)^{-1} I_{\phi,\psi}(\theta) I_{\psi,\psi}^{-1/2}(\theta) \right) \pi(\theta) d\theta$$

where $\mathbb{I}_p$ is the pxp identity matrix. Hence, $OI \in \mathbb{R}^{\dim(\psi) \mathrm{x} \dim(\psi)}$.

Variance Interpretation of RI

Consider two models. First, the **full model**, is where the parameter is (ψ, ϕ), with ψ interest, ϕ nuisance.

Variance Interpretation of RI

Consider two models. First, the **full model**, is where the parameter is (ψ, ϕ), with ψ interest, ϕ nuisance.

The asymptotic variance of the MLE $\hat{\psi}_{full}$ is then given by:

$$\lim_{n \to \infty} Var\left(\sqrt{n}\hat{\psi}_{full}\right) = I^{\psi,\psi}(\psi, \phi) = \left(I_{\psi,\psi} - I_{\psi,\phi} I_{\phi,\phi}^{-1} I_{\phi,\psi}\right)^{-1} \quad (5)$$

Variance Interpretation cont...

Now consider the **reduced model** where the nuisance parameters are considered to be known. In this model the only parameter is ψ.

Variance Interpretation cont...

Now consider the **reduced model** where the nuisance parameters are considered to be known. In this model the only parameter is ψ.

In this setting the asymptotic variance of the MLE $\hat{\psi}_{red}$ is given by:

$$\lim_{n\to\infty} Var\left(\sqrt{n}\hat{\psi}_{red} \mid \phi\right) = I_{red}^{\psi,\psi}(\psi,\phi) = \left(I_{\psi,\psi}\right)^{-1} \quad (6)$$

Note that this is also the asymptotic variance for an orthogonal parameterization.

Variance Interpretation cont...

The asymptotic relative efficiency (ARE) of the MLE in the joint case relative to the known (orthogonal) case is thus given by:

$$\lim_{n\to\infty} \frac{Var(\hat{\psi}_{red})}{Var(\hat{\psi}_{full})} = \frac{I_{\psi,\psi} - I_{\psi,\phi} I_{\phi,\phi}^{-1} I_{\phi,\psi}}{I_{\psi,\psi}} \tag{7}$$

Hence, providing some intuition behind RI and the OI.

Computation

- The index is simple to compute numerically by evaluating the information matrix over a grid of θ points.
- The Fisher Information, I, need only be computed in the original parameterization.
- User-supplied Jacobian matrix is only other requirement (& π)
- With symbolic computation only the transformation needs to be specified $\Rightarrow$ easy to try many different parameterizations
- π must satisfy $\int \pi d\theta < \infty$

Usage of OI

- First exhaust all other possibilities (i.e. both sets of PDE's)

Usage of OI

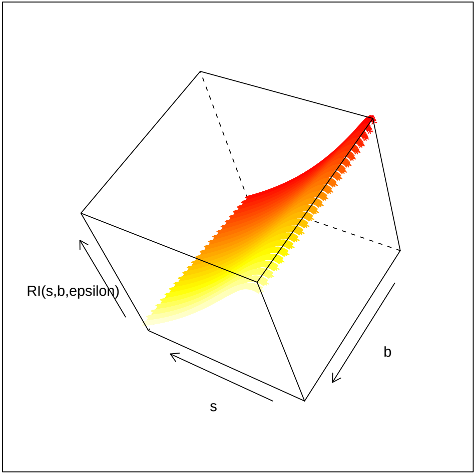
- First exhaust all other possibilities (i.e. both sets of PDE's)
- If unsuccessful, then considering searching for either:
- An 'approximately orthogonal' parameterization i.e. $OI \approx 1$ for general π, or,
- A 'locally orthogonal' parameterization i.e. $OI \approx 1$ for $\pi > 0$ only on some subset of Θ

Usage of OI

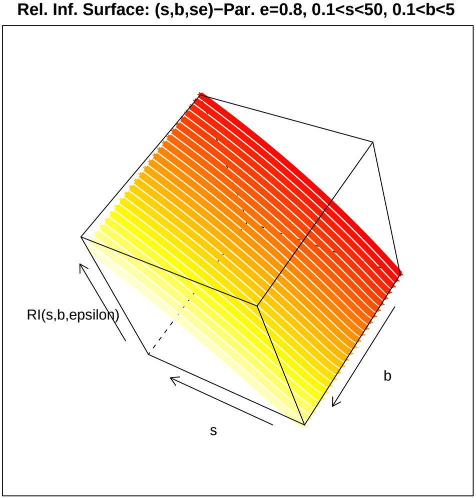
- First exhaust all other possibilities (i.e. both sets of PDE's)
- If unsuccessful, then considering searching for either:
- An 'approximately orthogonal' parameterization i.e. $OI \approx 1$ for general π, or,
- A 'locally orthogonal' parameterization i.e. $OI \approx 1$ for $\pi > 0$ only on some subset of Θ
- If this can be achieved then investigate coverage properties of the class of priors: $\pi(\theta) \propto \left(I^{11}\right)^{1/2} d(\theta_2)$

Example: (s, b, ϵ)−parameterization

Relative Information Surface: epsilon=0.8, 0.1<s<50, 0.1<b<5



EXAMPLE: $(s, \lambda_1, \lambda_2) = (s, b, s\epsilon)$−PARAMETERIZATION



Orthogonality

Note that if the parameterization is **orthogonal** then the first order PMP equation simplifies to:

$$\frac{\partial}{\partial \psi} \left\{ \pi(\psi, \phi) I_{\psi,\psi}^{-1/2} \right\} = 0 \tag{8}$$

The solution is seen to be:

$$\pi(\psi, \theta) = I_{\psi,\psi}^{1/2} \cdot d(\phi) \tag{9}$$

where $d(\phi)$ is an arbitrary smooth function of the nuisance parameter (Tibshirani, 1989).

Arbitrariness: good or bad? An interesting connection can help...

Reference Priors

Reference priors were first proposed by Bernardo (1979). Extended to **ordered group reference priors** in Berger & Bernardo (1989).

Reference Priors

Reference priors were first proposed by Bernardo (1979). Extended to **ordered group reference priors** in Berger & Bernardo (1989).

Idea: Divide parameters into groups of 'equal' (inferential) interest. $\theta_{(i)}$ is i^{th} most important of m groups. Generalization of interest/nuisance dichotomy.

Reference Priors

Reference priors were first proposed by Bernardo (1979). Extended to **ordered group reference priors** in Berger & Bernardo (1989).

Idea: Divide parameters into groups of 'equal' (inferential) interest. $\theta_{(i)}$ is i^{th} most important of m groups. Generalization of interest/nuisance dichotomy.

Let $m = 2$, $\theta_{(1)} = \psi$ be interest, with $\theta_{(2)} = \phi$ nuisance...

Berger-Bernardo Reference Prior Algorithm

1. Find the conditional reference prior for the nuisance parameter:
$$\pi(\phi|\psi) = |I_{\phi,\phi}(\psi,\phi)|^{1/2}$$
2. Typically this is improper. Choose a sequence of subsets of the parameter space $\Omega_{i,\psi}$ over which to normalize:
$$p_i(\phi|\psi) = \pi(\phi|\psi) \cdot K_i(\psi) \cdot \mathbf{1}_{\psi \in \Omega_{i,\psi}}$$
where:
$$K_i(\psi) = \left[\int_{\Omega_{i,\psi}} \pi(\phi|\psi) d\phi \right]^{-1}$$

B-B Algorithm cont. . .

3. Find the marginal reference prior for ψ wrt $p_i(\phi|\psi)$:

$$\pi_i(b, \epsilon) = \exp\left\{ \frac{1}{2} \int_{\Omega_{i,\psi}} p_i(\phi|\psi) \cdot \log\left[\frac{|I(\psi,\phi)|}{|I_{\phi,\phi}(\psi,\phi)|} \right] d\phi \right\}$$

4. Finally, the reference prior is defined to be:

$$\pi(\psi, \phi) = \lim_{i \to \infty} \left[\frac{K_i(\psi)\pi_i(\psi)}{K_i(\psi_0)\pi_i(\psi_0)} \right] \pi(\phi|\psi)$$

where ψ_0 is any fixed point within the chosen compact subsets.

Properties/Connections

Important things to note:

Def. The 'reverse reference prior' (RRP) switches roles of ψ, ϕ

- **Under an orthogonal parameterization** the RRP is first order probability matching (Berger [via J.K.Ghosh], 1992)
- $\psi-$dependence is determined entirely though:

$$\pi(\psi|\phi) \propto |I_{\psi,\psi}(\psi,\phi)|^{1/2}$$

So it is just of the Tibshirani class!

- PM property not guaranteed outside orthogonality, but prior still derived from sound information-theoretic principles.

Special Cases of PMPs

(1) In the univariate case ($p = 1$), Jeffreys prior is the *unique* PMP!

(2) Jeffreys prior is NOT necessarily probability matching for $p > 1$

(3) For orthogonal settings, the RRP is first order matching

(4) The regular reference prior need not be (but often is)

(5) In some cases, it can be proved that there is no PMP!

(6) Two-sided intervals are first order PM for *any* prior (Hartigan, 1966)

Why not just orthogonalize?

- So, we have a nice class of priors (Tibshirani), with the RRP one potentially appealing case within this class. They will be PMPs under orthogonality...

Why not just orthogonalize?

- So, we have a nice class of priors (Tibshirani), with the RRP one potentially appealing case within this class. They will be PMPs under orthogonality...

- Datta & Ghosh (1996) and Mukerjee & Ghosh (1997) noted the invariance of matching priors under reparametrization

Why not just orthogonalize?

- So, we have a nice class of priors (Tibshirani), with the RRP one potentially appealing case within this class. They will be PMPs under orthogonality. . .
- Datta & Ghosh (1996) and Mukerjee & Ghosh (1997) noted the invariance of matching priors under reparametrization
- Cox & Reid (1987) showed that, in theory, we can always orthogonalize a univariate interest parameter and a $(p-1)-$dimensional nuisance parameter

Why not just orthogonalize?

- So, we have a nice class of priors (Tibshirani), with the RRP one potentially appealing case within this class. They will be PMPs under orthogonality...
- Datta & Ghosh (1996) and Mukerjee & Ghosh (1997) noted the invariance of matching priors under reparametrization
- Cox & Reid (1987) showed that, in theory, we can always orthogonalize a univariate interest parameter and a $(p-1)-$dimensional nuisance parameter
- Unfortunately, in practice, this is often not feasible as a set of $p-1$ PDE's must be solved...

The Problem in a Nutshell

Both obvious routes to finding probability matching priors:

1. Directly from the characterization theorem (3), or,
2. Via orthogonal parameterization

are blocked by the obstacle of an intractable (set of) PDE('s)!

The Problem in a Nutshell

Both obvious routes to finding probability matching priors:

1. Directly from the characterization theorem (3), or,
2. Via orthogonal parameterization

are blocked by the obstacle of an intractable (set of) PDE('s)!

The third possible route – (Reverse) Reference priors – is also often overwhelmingly complicated to compute!

"the theory of Bayesian objectivity cannot be a simple one"

Efron (1986), quoted in Berger & Bernardo (1992).

Applications

LHC Example

Recall the three-Poisson example earlier. In this case the PMP equation is:

$$\frac{\partial}{\partial s}\left\{\pi(s,b,\epsilon)\sqrt{\frac{b}{\epsilon t u}\left[\frac{st(u+s)}{b}+\frac{u(1+t)}{\epsilon}\right]}\right\}+$$
$$\frac{\partial}{\partial b}\left\{-\frac{b\cdot\pi(s,b,\epsilon)}{\epsilon t}\left(\frac{b}{\epsilon t u}\left[\frac{st(u+s)}{b}+\frac{u(1+t)}{\epsilon}\right]\right)^{-1/2}\right\}+$$
$$\frac{\partial}{\partial \epsilon}\left\{-\frac{s\cdot\pi(s,b,\epsilon)}{u}\left(\frac{b}{\epsilon t u}\left[\frac{st(u+s)}{b}+\frac{u(1+t)}{\epsilon}\right]\right)^{-1/2}\right\}= \quad 0$$

Computational Difficulties

The previous equation has so far proved too complex to solve even using Mathematica, Maple etc. The multi-channel is even more daunting...

The regular reference prior is also brutal to compute (limits in $4M^2$ competing directions and $4M$ integrations). However, the RRP frequently has better matching properties and has yielded some luck...

Any Hope?

Conjecture

My Conjecture: *There exists no PMP for this particular example.*

Proof.

No formal proof. . . hence it is just a conjecture! □

RRP Computation

Derivation for general $M-$channel setting:

1. Conditional reference prior:

$$\pi(s|\mathbf{b},\epsilon) \propto \sqrt{\sum_{j=1}^{M} \frac{\epsilon_j^2}{s\epsilon_j + b_j}}$$

2. Normalizing constant on $(s_{(l_i)}, s_{(u_i)})$:

$$K_i(\mathbf{b},\epsilon) = s_{(u_i)}^{1/2}\left[2\sqrt{\sum_{j=1}^{M}\epsilon_j}\right] - s_{(l_i)}\sqrt{\sum_{j=1}^{M}\frac{\epsilon_j^2}{b_j}} + O\left(s_{(u_i)}^{-1/2}\right) + O\left(s_{(l_i)}^2\right)$$

RRP Computation cont...

3. The marginal prior:

$$\propto \exp\left\{ \frac{s_{(u_j)}^{1/2}\left[\sqrt{\sum \epsilon_j}\right] \cdot \left(2\log s_{(u_j)} + \log\left[\frac{\left(\prod t_j\right)\sum \epsilon_j u_j}{\left(\prod b_j \epsilon_j\right)\sum \epsilon_j}\right] - 4\right) + \sqrt{O\left(s_{(u_j)}^{-1}\right) + O(s^{(}l_j))}}{s_{(u_j)}^{1/2}\left[2\sqrt{\sum \epsilon_j}\right] + O\left(s_{(u_j)}^{-1/2}\right) + O\left(s_{(l_j)^2}\right)} \right\}$$

4. The limit can be shown to yield the RRP:

$$\pi(s,b,\epsilon) \propto \sqrt{\sum_{j=1}^{M} \frac{\epsilon_j^2}{s\epsilon_j + b_j}} \cdot \frac{1}{\sum_{j=1}^{M} \epsilon_j} \cdot \sqrt{\frac{\sum_{j=1}^{M} \epsilon_j u_j}{\prod_{j=1}^{M} b_j \epsilon_j}}$$

The Regular Reference Prior

The regular reference prior for the ordered parameterization $(\psi = s, \phi = (\mathbf{b}, \epsilon))$, if it exists, will be of the form:

$$\pi(s, \mathbf{b}, \epsilon) \propto g(s) \sqrt{\prod_{j=1}^{M} \frac{b_j u_j (1 + t_j) + \epsilon_j s t_j (s + u_j)}{b_j \epsilon_j (b_j + \epsilon_j s)}} \qquad (10)$$

Where $g(\cdot)$ is a smooth function of s alone (that could, in principle, be determined by complicated limit calculations). Heuristics suggest that $g(s) \approx s^{-\delta}$ with $\delta > 1$ although this is not rigorous!

Relating to the conjecture

For the single-channel setting ($M = 1$):

1. The reference prior **cannot** be a PMP!
   [for the ordered parameterization $(\psi = s, \phi = (\mathbf{b}, \epsilon))$]
2. Priors of the Tibshirani class $\pi(s, \mathbf{b}, \epsilon)$ **cannot** be PMPs!
3. Hence, the reverse reference prior **cannot** be a PMP!
   [for the ordered parameterization $(\psi = s, \phi = (\mathbf{b}, \epsilon))$]

Prospects look grim for standard priors. May wish to consider data-dependent priors as a mathematical tool.

Simulation Study

Details:

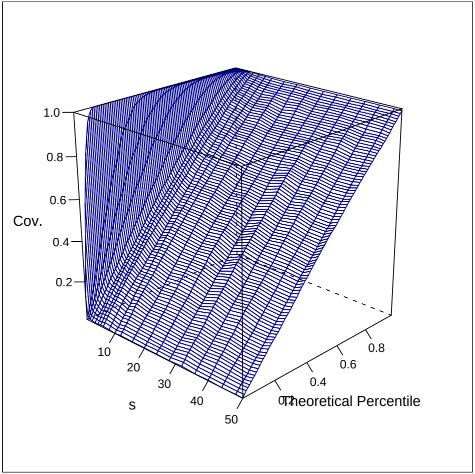
- 110,000 datasets generated, corresponding to 22 different s values: 0.1 to 50.0
- Fixed $\epsilon = 1$, $b = 3$
- Coverage properties computed for each percentile: $\{s^{(0.01)}, \ldots, s^{(0.99)}\}$.

Compare performance based on coverage surfaces (Goal: 45° plane)...

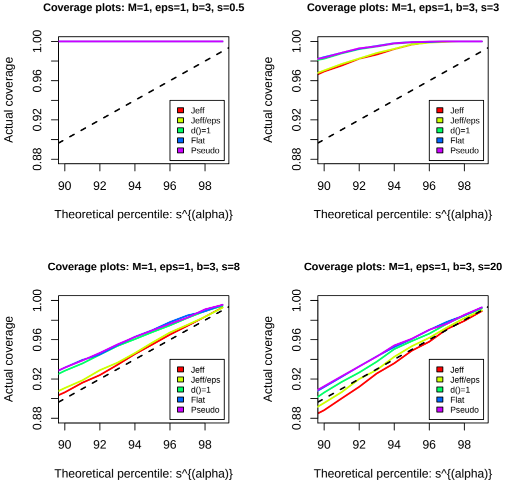
Single-channel results Ten-channel results

Results

Coverage surface for d()=1 prior: e=1,b=3



Results


Coverage plots: M=1, eps=1, b=3, s=0.5
Coverage plots: M=1, eps=1, b=3, s=3
Coverage plots: M=1, eps=1, b=3, s=8
Coverage plots: M=1, eps=1, b=3, s=20
Actual coverage
0.88
0.92
0.96
1.00
90
92
94
96
98
Theoretical percentile: s^{(alpha)}
Jeff
Jeff/eps
d()=1
Flat
Pseudo

Summary Results Format

(Simulated data specification) $M = 1 : (b = 3, \epsilon = 1, s = 0.5)$
Fictional Coverage table (actual coverage of the percentiles):

Percentile	Prior 1	Prior 2	Prior 3	Prior 4	Prior 5
$s^{(0.05)}$	0.05	0.10	0.01	0.16	0.08
$s^{(0.10)}$	0.10	0.25	0.03	0.33	0.15
$s^{(0.25)}$	0.25	0.75	0.21	0.51	0.28
$s^{(0.50)}$	0.50	0.95	0.40	0.55	0.51
$s^{(0.75)}$	0.75	1.00	0.62	0.64	0.77
$s^{(0.90)}$	0.90	1.00	0.80	0.76	0.91
$s^{(0.95)}$	0.95	1.00	0.88	0.80	0.95
$s^{(0.99)}$	0.99	1.00	0.92	0.90	0.98
	Perfect!	Overcover	Undercover	Over&Under	Typical

Results: Original parameterization

$M = 1 : (b = 3, \epsilon = 1, s = 0.5)$

Coverage table:

	Jeff	RRP	Jeff/ϵ	$d = 1$	Flat	Pseudo
$s^{(0.05)}$	0.11	0.15	0.11	0.15	0.16	0.15
$s^{(0.10)}$	0.25	0.33	0.25	0.33	0.33	0.33
$s^{(0.25)}$	0.70	0.82	0.71	0.81	0.81	0.81
$s^{(0.50)}$	1.00	1.00	1.00	1.00	1.00	1.00
$s^{(0.75)}$	1.00	1.00	1.00	1.00	1.00	1.00
$s^{(0.90)}$	1.00	1.00	1.00	1.00	1.00	1.00
$s^{(0.95)}$	1.00	1.00	1.00	1.00	1.00	1.00
$s^{(0.99)}$	1.00	1.00	1.00	1.00	1.00	1.00

Simulation Results: $M = 1,\ s = 8$

$M = 1 : (b = 3, \epsilon = 1, s = 8)$

Coverage table:

	Jeff	RRP	Jeff/ϵ	$d = 1$	Flat	Pseudo
$s^{(0.05)}$	0.05	0.07	0.05	0.06	0.07	0.07
$s^{(0.10)}$	0.10	0.13	0.11	0.13	0.13	0.13
$s^{(0.25)}$	0.24	0.29	0.25	0.28	0.29	0.29
$s^{(0.50)}$	0.49	0.55	0.50	0.54	0.55	0.55
$s^{(0.75)}$	0.75	0.80	0.76	0.79	0.80	0.80
$s^{(0.90)}$	0.91	0.93	0.91	0.93	0.93	0.93
$s^{(0.95)}$	0.96	0.97	0.96	0.97	0.97	0.97
$s^{(0.99)}$	0.99	1.00	0.99	0.99	1.00	1.00

Simulation Results: $M = 1,\ s = 50$

$M = 1 : (b = 3, \epsilon = 1, s = 50)$

Coverage Table:

	Jeff	RRP	Jeff/ϵ	$d = 1$	Flat	Pseudo
$s^{(0.05)}$	0.05	0.06	0.06	0.05	0.06	0.06
$s^{(0.10)}$	0.10	0.12	0.11	0.11	0.12	0.12
$s^{(0.25)}$	0.25	0.28	0.28	0.26	0.28	0.28
$s^{(0.50)}$	0.52	0.55	0.54	0.52	0.55	0.55
$s^{(0.75)}$	0.76	0.78	0.77	0.77	0.78	0.78
$s^{(0.90)}$	0.90	0.91	0.91	0.90	0.91	0.91
$s^{(0.95)}$	0.95	0.96	0.95	0.95	0.96	0.96
$s^{(0.99)}$	0.99	0.99	0.99	0.99	0.99	0.99

$M = 10$-Channel Example

Now, we return to the multi-channel example with $M = 10$.

$M = 10$-Channel Example

Now, we return to the multi-channel example with $M = 10$.

(1) **Flat:** $\pi_1(s, b, \epsilon) \propto 1$

(2) **Jeffreys:** $\pi_2(s, b, \epsilon) \propto \sqrt{\det(I(s, b, \epsilon))}$

(3) **Reverse Reference:**

$$\pi_3(s, b, \epsilon) \propto \sqrt{\sum_{j=1}^{M} \frac{\epsilon_j^2}{s\epsilon_j + b_j}} \cdot \frac{1}{\sum_{j=1}^{M} \epsilon_j} \cdot \sqrt{\frac{\sum_{j=1}^{M} \epsilon_j u_j}{\prod_{j=1}^{M} b_j \epsilon_j}}$$

(4) $\pi_4(s, b, \epsilon) \propto \sqrt{I_{ss}(s, b, \epsilon)}$

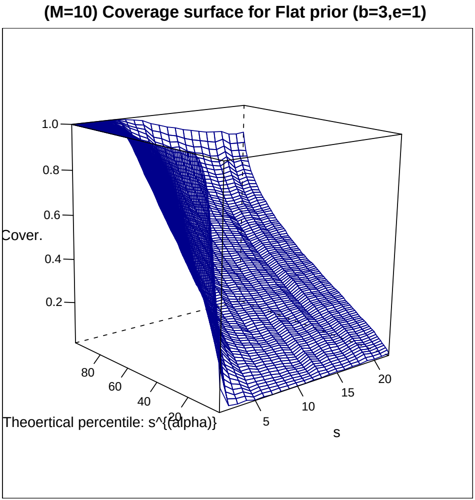
(5) $\pi_5(s, b, \epsilon) \propto \sqrt{I_{ss}(s, b, \epsilon)} \frac{1}{\sqrt{\epsilon_1 \cdots \epsilon_M}}$

(6) $\pi_6(s, b, \epsilon) \propto \sqrt{I_{ss}(s, b, \epsilon)} \frac{1}{\epsilon_1 \cdots \epsilon_M}$

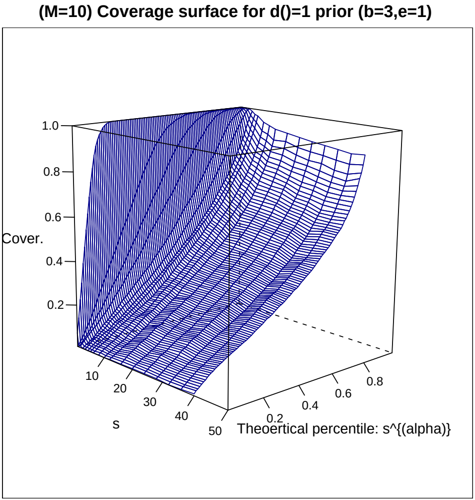
(7) $\pi_7(s, b, \epsilon) \propto \sqrt{I_{ss}(s, b, \epsilon)} \frac{1}{\epsilon_1 \cdots \epsilon_M \cdot b_1 \cdots b_M}$

Again, compare over 22 values of s, with $b_i \sim N(0.3, 0.04^2)$, $\epsilon_i \sim N(0.1, 0.025^2)$.

Performance: Ten-Channel Results



$d(\cdot) = 1$ prior: Ineffective!



Numerical Results: Ten-Channel

$M = 10 : (b = 3, \epsilon = 1, s = 0.5)$

Coverage Table:

	Flat	Jeff	RRP	$d = 1$	$d = \frac{1}{\sqrt{\epsilon}}$	$d = \frac{1}{\epsilon}$	$d = \frac{1}{\epsilon \cdot b}$
$s^{(0.05)}$	0.11	0.08	0.10	0.08	0.13	0.09	0.10
$s^{(0.10)}$	0.22	0.18	0.21	0.17	0.26	0.20	0.22
$s^{(0.25)}$	0.53	0.51	0.52	0.48	0.60	0.53	0.55
$s^{(0.50)}$	0.93	0.94	0.94	0.92	0.97	0.95	0.97
$s^{(0.75)}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$s^{(0.90)}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$s^{(0.95)}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00
$s^{(0.99)}$	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Numerical Results: Ten-Channel

$M = 10 : (b = 3, \epsilon = 1, s = 10)$

Coverage table:

	Flat	Jeff	RRP	$d = 1$	$d = \frac{1}{\sqrt{\epsilon}}$	$d = \frac{1}{\epsilon}$	$d = \frac{1}{\epsilon \cdot b}$
$s^{(0.05)}$	0.06	0.04	0.05	0.03	0.04	0.06	0.07
$s^{(0.10)}$	0.11	0.07	0.09	0.05	0.07	0.11	0.13
$s^{(0.25)}$	0.23	0.19	0.20	0.15	0.18	0.23	0.26
$s^{(0.50)}$	0.41	0.37	0.38	0.31	0.37	0.41	0.45
$s^{(0.75)}$	0.63	0.58	0.60	0.53	0.58	0.63	0.66
$s^{(0.90)}$	0.79	0.76	0.78	0.72	0.77	0.80	0.82
$s^{(0.95)}$	0.87	0.84	0.85	0.81	0.85	0.87	0.89
$s^{(0.99)}$	0.96	0.95	0.95	0.93	0.95	0.96	0.97

Ten-Channel Summary

(1) Much harder than the one-channel case

(2) Far more important to appropriately select the $d(\cdot)$ function

(3) Improved choices of $d(\cdot)$ are available but further simulation studies required

Future Work

This is certainly a topic with much room for development, and (hopefully) rich rewards...

- Analytic approximations to PMP ('closest' PMP: metric?)
- Utilize the huge literature on asymptotic statistics...
- Reduce computational burden/improve efficiency
- Explore deep connections with other aspects of asymptotic/Bayesian theory e.g. SOUP (Meng & Zaslavsky, 2002)
- Extensive simulation studies
- **'The Holy Grail of PMP'**: A general framework to implement first and second order PMP's (unlikely anytime soon...)

Conclusion

In summary:

Conclusion

In summary:

1. (Where they exist...) PMP's may offer an 'optimal solution'... (...depending on the criteria...)
2. No PMP $\nRightarrow$ No good Bayesian inference! Other criteria...
3. Computational challenges for PMPs yet to be overcome in the general case (much work to be done!)
4. PMP's are simple to obtain in orthogonal settings... (...but are somewhat arbitrary)
5. Even reference priors, usually considered the 'gold standard' in default priors, struggle to provide an entirely satisfactory solution
6. May apply PMP's from the orthogonal setting in 'almost orthogonal' parameterizations
7. *OI* score to determine 'how orthogonal' a parameterization is

References

1. Berger, J.O., Bernardo, J.M. (1992) Ordered Group Reference Priors with Application to the Multinomial Problem. *Biometrika* **79**, 25-37
2. Bernardo, J.M. (1979) Reference posterior distributions for Bayesian inference (with discussion). *Journal of the Royal Statistical Society, B* **41**, 113-147
3. Clarke, B., Wasserman, L. (1993) Noninformative Priors and Nuisance Parameters. *Journal of the American Statistical Association* **88**, 1427-1432
4. Cox, D.R. & Reid, N. (1987) Parameter orthogonality and approximate conditional inference (with discussion). *Journal of the Royal Statistical Society, B* **49**, 1-39
5. Datta, G.S., and Mukerjee, R. (2004) Probability Matching Priors: Higher Order Asymptotics (Lecture Notes in Statistics), *Springer* **178**
6. Demortier, L. (2005) Bayesian Reference Analysis. *Proceedings of PHYSTAT 2005*, `http://www.physics.ox.ac.uk/phystat05/proceedings/files/demortier-refana.ps`
7. Hartigan, J.A. (1966) Note on the Confidence-Prior of Welch and Peers. *Journal of the Royal Statistical Society, Series B* **28** 55-56

References

8. Kass, R.E., Wasserman, L. (199?) The Selection of Prior Distributions by Formal Rules. *Journal of the American Statistical Association* **91** 1343-1370
9. Levine, R.A. & Casella, G. (2003) Implementing probability matching priors for Frequentist inference. *Biometrika* **90**, 127-137
10. Meng, X.-L., Zaslavsky, A.M. (2002) Single observation unbiased priors. *Ann. Statist.* **30-5** 1345-1375.
11. Mukerjee, R. & Ghosh, M. (1997) Second-order probability matching priors. *Biometrika* **86**, 333-340.
12. Novick, M.R., and Hall, W.J. (1965), A Bayesian Indifference Procedure, *Journal of the American Statistical Association*, **60** 1104-1117
13. Peers, H.W. (1965) On Confidence Points and Bayesian Probability Points in the Case of Several Parameters. *Journal of the Royal Statistical Society, Series B* **27** 9-16
12. Sweeting, T.J. (2005) On the implementation of local probability matching priors for interest parameters. *Biometrika* **92**, 47-57
13. Tibshirani, R.J. (1989) Noninformative priors for one parameter of many. *Biometrika* **76**, 604-608
14. Zellner, A. (1971), An Introduction to Bayesian Inference in Econometrics, *New York: John Wiley*