Probabilistic inference without (really) using Bayes[1]

Ryan Martin
North Carolina State University
www4.stat.ncsu.edu/~rmartin

Statistics Seminar
North Carolina State University
12/02/2016

[1]Some of this work is/was supported by grants from NSF and U.S. Army.

Introduction

- *Statistical inference is the process of converting data, prior information, modeling assumptions, etc, into a meaningful probability-like summary of uncertainty about the true state of the system under investigation.*

- Therefore, my goal is always to obtain a meaningful posterior probability distribution — *or something similar* — for the unknown quantities of interest.

Introduction, cont.

- This process can be implemented via a Bayesian approach, but there are several potential shortcomings.
- One in particular: *Bayes requires a model for everything*.
- But we often don't have enough information to justify a choice of model at all levels.
  - "Bad" choices can obviously mess things up.
  - Being completely robust/nonparametric can be overkill.
  - Model uncertainty can be difficult to account for.
- Therefore, some deviation from the standard Bayes approach seems reasonable, maybe even necessary.
- My main research focus is in exploring various deviations from Bayes and in developing some alternative approaches.

This talk

Discuss three deviations from Bayes, with some applications, results, and open questions.

Case 1. *Partial prior information*.
Sparsity assumptions in high-dim problems.

Case 2. *"Misspecification on purpose."*
No likelihood, or don't want to use it.

Case 3. *Probabilistic inference without Bayes*.
When we don't have a model for everything, is Bayes (or probability) even appropriate?

Case 1: partial prior information

- Canonical normal mean model: $Y_1, \ldots, Y_n$ independent with $Y_i \sim \mathrm{N}(\theta_i, 1)$, $i = 1, \ldots, n$, with n large.
- High-dimensional — each Y_i has its own parameter θ_i.
- *Sparsity:* most of the θ_i's are zero.
- Sparsity acts like partial prior information.
- We don't have anything else to help us formulate a full prior for θ, but Bayes still says we need one...

Case 1, cont.

- Standard approach is to express θ as a pair (S, θ_S):
  - $S \subseteq \{1, 2, \ldots, n\}$ denotes the location of non-zeros
  - θ_S the $|S|$-vector of non-zero values.
- Sparsity helps to write down a meaningful prior for S
- What about the conditional prior for θ_S, given S?
  - Normal: computationally good but theoretically not-so-good;
  - Laplace: theoretically good but computationally not-so-good.
- This is silly — both priors are meaningless!
- Why not make another (possibly equally meaningless) choice that doesn't sacrifice on either theory or computation?
- *Idea:* use data in the "prior" for θ_S given S...

Case 1, cont.

- Laplace is theoretically good because it's heavy-tailed.
- But tails don't matter if the prior is properly centered.
- Use data to center the computationally good normal prior.
- In particular, take conditional "prior" as

$$(\theta_S \mid S) \sim \mathsf{N}_{|S|}(Y_S, \gamma^{-1} I_{|S|}), \quad \gamma \in (0, 1).$$

- Conditional prior for θ_S, given S, times the marginal prior for S gives a data-dependent prior Π_n for θ.

Case 1, cont.

- Turns out that this simple data-dependent prior is too greedy.
- To correct for this, do a regularization step:

$$\widetilde{\Pi}_n(d\theta) := \frac{\Pi_n(d\theta)}{L_n(\theta)^{1-\alpha}}, \quad \alpha \in (0,1),$$

  where $L_n(\theta)$ is the likelihood function.
- *Double empirical Bayes*

$$\Pi^n(d\theta) \propto \begin{cases} L_n(\theta)\,\widetilde{\Pi}_n(d\theta), \text{ or equivalently} \\ L_n(\theta)^{\alpha}\,\Pi_n(d\theta). \end{cases}$$

- Double = Centering + Regularization.
- Is the posterior Π^n reasonable?

Case 1, cont.

- "Best of both worlds:"
  - Computationally convenient because it's a normal prior.
  - Theoretically, it also has the optimal concentration rate.
- For the theory:
  - $\Theta_n = \{\theta \in \mathbb{R}^n : \|\theta\|_0 = s_n\}$, where $s_n = o(n)$.
  - $\varepsilon_n = s_n \log(n/s_n)$ is the minimax optimal rate for Θ_n.

Posterior concentration theorem.

For suitable sparsity prior on S, for any $(\alpha, \gamma) \in (0,1)^2$, there exists $M > 0$ such that

$$\sup_{\theta^\star \in \Theta_n} \mathsf{E}_{\theta^\star} \Pi^n(\{\theta \in \mathbb{R}^n : \|\theta - \theta^\star\|_2^2 > M\varepsilon_n\}) \to 0, \quad n \to \infty.$$

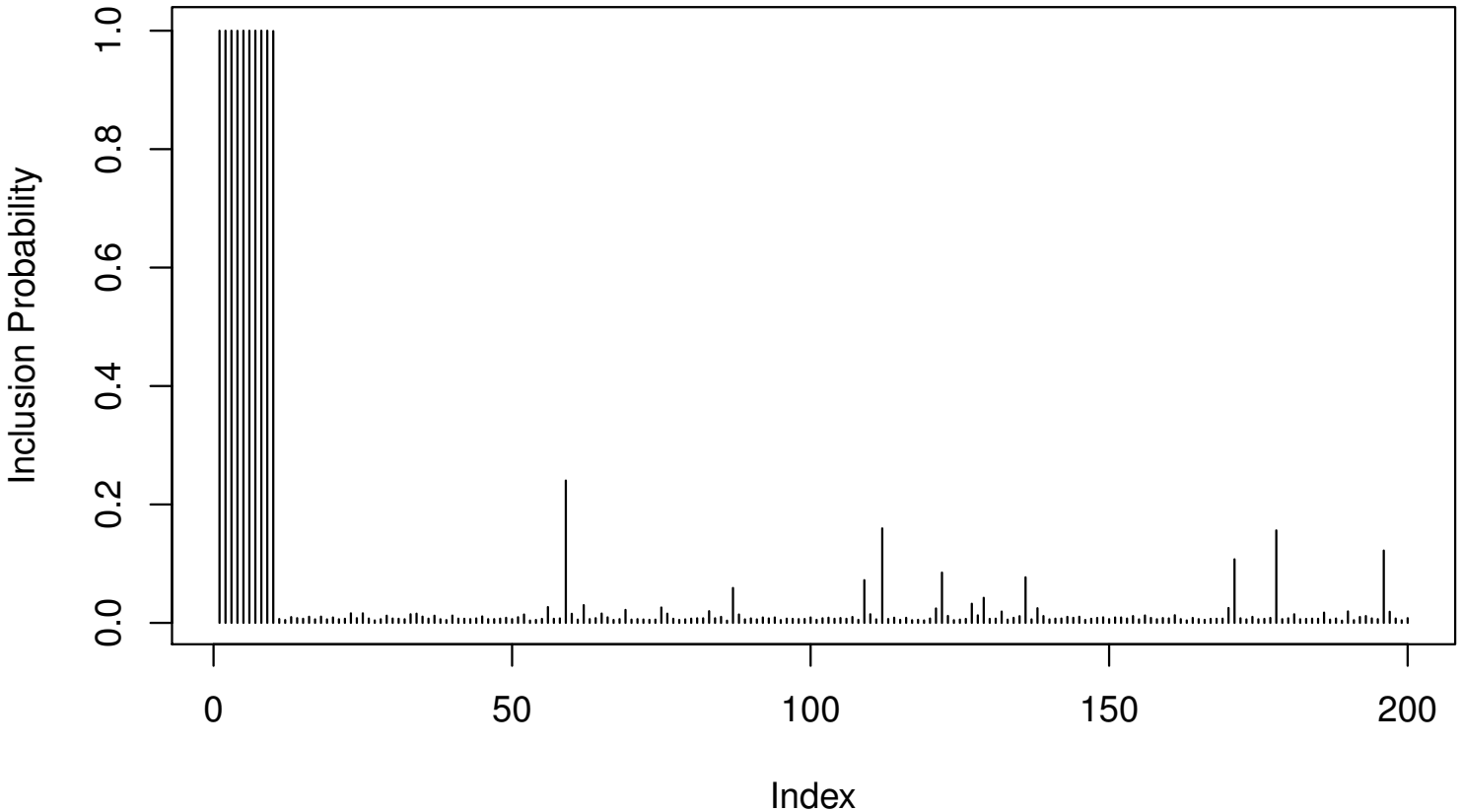
Case 1, cont.

Mean square error loss comparisons. $n = 200$, s_n is effective dimension, constant signals of size A; first three rows come from the Dirichlet–Laplace paper (*JASA* 2015).

s_n	10		20		40	
A	7	8	7	8	7	8
$DL_{1/n}$	16	14	33	31	66	60
EBMed	26	26	57	56	119	119
PMed1	23	22	49	48	102	102
DEB*	**13**	**13**	**25**	**25**	**47**	**48**

(*using $\alpha = 0.99$ and $\gamma = 0.01$.)

Case 1, cont.



$n = 200$, $s_n = 10$, and $\theta_1^\star = \cdots = \theta_{10}^\star = 7$, others zero.

Case 1: Remarks

- Other kinds of concentration rate results can be proved, e.g., effective dimension of Π^n.
- DEB for high-dim linear regression has been done, stronger rate results than lasso-based methods.
- Other high-dim problems admit a representation of the form (S, θ_S), e.g., density estimation with mixtures; we have some general rate results for these problems too.
- Some open questions...
  - Basically only worked out theory so far, can we build some useful methodology?
  - Theory and methods for some other interesting high-dim problems, e.g., GLMs, graphical models, ...?
  - Coverage of credible regions?

Case 1: References

- M. and Walker (2014). Asymptotically minimax empirical Bayes estimation of a sparse normal mean vector. *Electronic Journal of Statistics*.
- M., Mess, and Walker (2016+). Empirical Bayes posterior concentration in sparse high-dimensional linear models. *Bernoulli*.
- M. and Walker (2016+). Optimal Bayesian posterior concentration rates with empirical priors. `arXiv:1604.05734`.

Case 2: "misspecification on purpose"

- *Motivation:* iid sample $Y_1, \ldots, Y_n$ from a distribution and the goal is inference on the median θ.
- A non-Bayesian would have no trouble with this!
- A Bayesian apparently needs to introduce a likelihood.
  - A "bad" likelihood might mess up inference on θ.
  - A "good" likelihood may involve other parameters, a modeling and computational nuisance.
- I like having a posterior, but would be great if I could get it directly, without introducing a likelihood...

Case 2, cont.

- Consider a *Gibbs model*.
- Write $R(\theta) = \mathrm{E}|Y_1 - \theta|$; true median $\theta^\star$ minimizes $R(\cdot)$.
- Get a prior Π for θ, and compute Gibbs posterior

$$\Pi^n(d\theta) \propto e^{-\omega n R_n(\theta)}\, \Pi(d\theta).$$

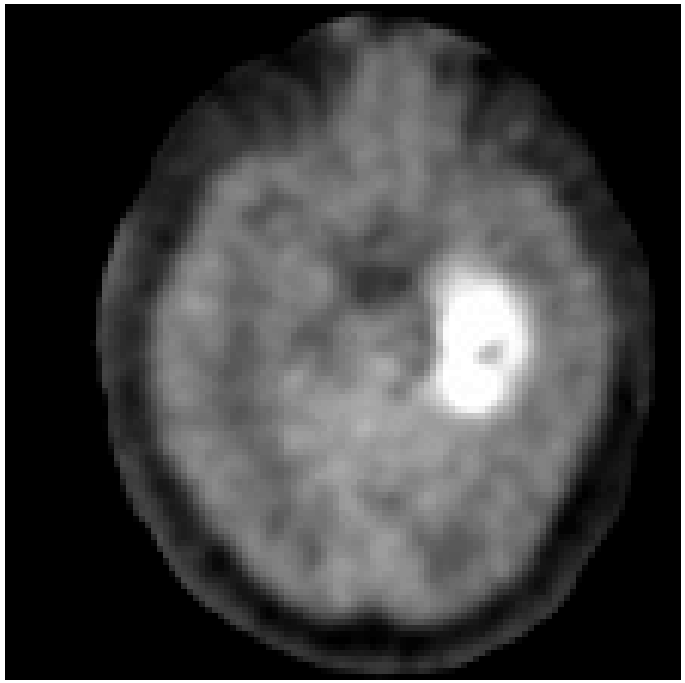
  - $R_n(\theta) = n^{-1} \sum_{i=1}^n |Y_i - \theta|$, empirical version of $R(\theta)$;
  - ω is a scalar tuning parameter.
- Basically a posterior based on *purposely misspecified* model.
- Highlights:
  - Has desirable concentration rate properties...
  - Basically no nuisance parameters — prior specification and posterior computation only for interest parameter!
  - Scale ω is important, controls spread/calibration...

Case 2, cont.

- Nothing special about the median!
- Whenever true $\theta^\star$ is the minimizer of a risk $R(\theta)$, a Gibbs model can/should be used.
  - Sometimes R is given,
  - other times we need to cook it up ourselves.
- Some specific applications we've looked at so far:
  - minimum clinically important difference
  - quantile regression
  - image boundary detection

Case 2, cont.

- Data (X_i, Y_i) are pixel locations and intensities, $i = 1, \ldots, n$.
- Intensity measurements tend to be stronger inside an unknown region Γ compared to outside.
- Goal is to make inference on Γ...



Case 2, cont.

- A fully Bayesian model is possible and has been done.
- Requires modeling intensities:
  - A "good" model won't help inference on Γ
  - but a "bad" model might hurt...
- Can we construct a Gibbs posterior for Γ?
- Trick is defining $R(\Gamma)$ so that true $\Gamma^\star$ satisfies

$$R(\Gamma) > R(\Gamma^\star), \quad \forall\, \Gamma \neq \Gamma^\star.$$

- *Roughly:* we showed that a twist on missclassification error, one that has scale ω built in, will do the job.

Case 2, cont.

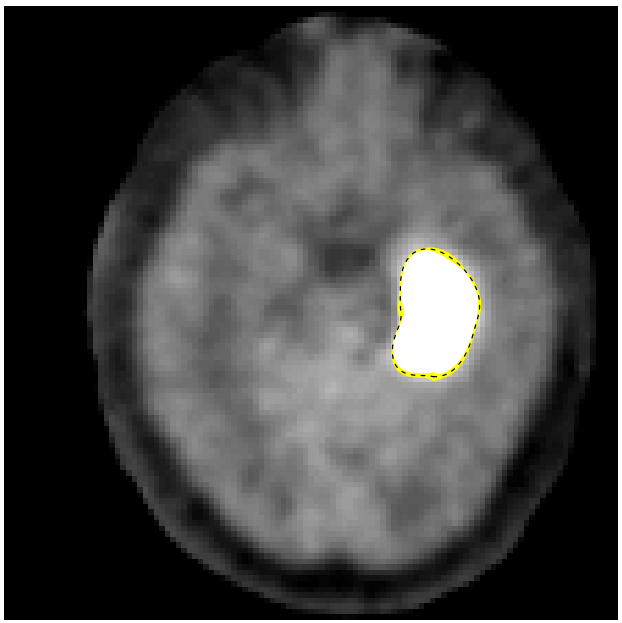
- Characterize Γ by its boundary $\gamma = \partial\Gamma$.
- Treat γ as a function, prior is a mixture of b-splines.
- For theory:
  - True $\Gamma^\star$ in class $\mathcal{H}(\alpha)$ with α-Hölder smooth boundary.
  - Optimal rate on $\mathcal{H}(\alpha)$ is $\varepsilon_n = \{\log(n)/n\}^{\alpha/(\alpha+1)}$.
  - Prior doesn't know $\alpha = \alpha(\Gamma^\star)$, so our rate is adaptive.

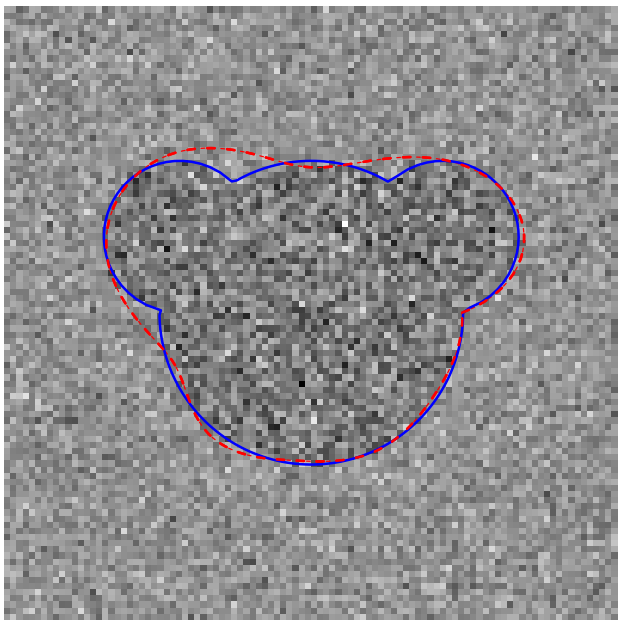
Posterior concentration theorem.

For b-spline mixture prior, suitable risk $R(\Gamma)$, and any $M_n \to \infty$,

$$\sup_{\Gamma^\star \in \mathcal{H}(\alpha)} \mathsf{E}_{\Gamma^\star} \Pi^n(\{\Gamma : \lambda(\Gamma \triangle \Gamma^\star) > M_n \varepsilon_n\}) \to 0, \quad n \to \infty.$$

Case 2, cont.





Case 2: Remarks

- Direct attack on interest parameter, minimal prior specification and posterior computation requirements.
- Robust because it's likelihood-free, can avoid introducing high- or infinite-dim nuisance parameters.
- Some cases, e.g., general Lévy process models, there is a likelihood but it can't be written down.
- Some work on choice of ω has been done...
- Open questions:
  - More applications? Finding $R(\theta)$ is tricky...
  - More efficient methods for scaling with ω...?

Case 2: References

- Syring and M. (2016+). Gibbs posterior inference on the minimum clinically important difference. `arXiv:1501.01840`.
- Syring and M. (2016+). Calibrating general posterior credible regions. `arXiv:1509.00922`.
- Syring and M. (2016+). Robust and rate-optimal Gibbs posterior inference on the boundary of a noisy image. `arXiv:1606.08400`.
- M., Ouyang, and Domagni (2016+). Efficient posterior inference on the volatility of a jump diffusion process. `arXiv:1608.06663`.

Case 3: probabilistic inference without Bayes

- Consider the classical textbook inference problem:
  - Given (parametric) model, $Y \sim \mathrm{P}_\theta$;
  - No prior information for θ;
  - Goal is inference on θ.
- Probabilistic inference without a prior seems out of reach.
- Brad Efron has called it the
  - "Holy Grail"
  - "most important unresolved problem in statistics"
- Naturally, many (including Fisher) have tried but...

Case 3, cont.

- *What makes the problem so difficult?*
- Meaningful probabilities don't come from thin air!
- I can write down any posterior probabilities that I want, but how do I know if they're meaningful?
- *Vague Claim 1:*

  probabilities are only meaningful relative to the model that defines them.

- Our inference problem is missing exactly what's needed to make probabilities meaningful.
- *Is probability even the right tool in this context?* [2]

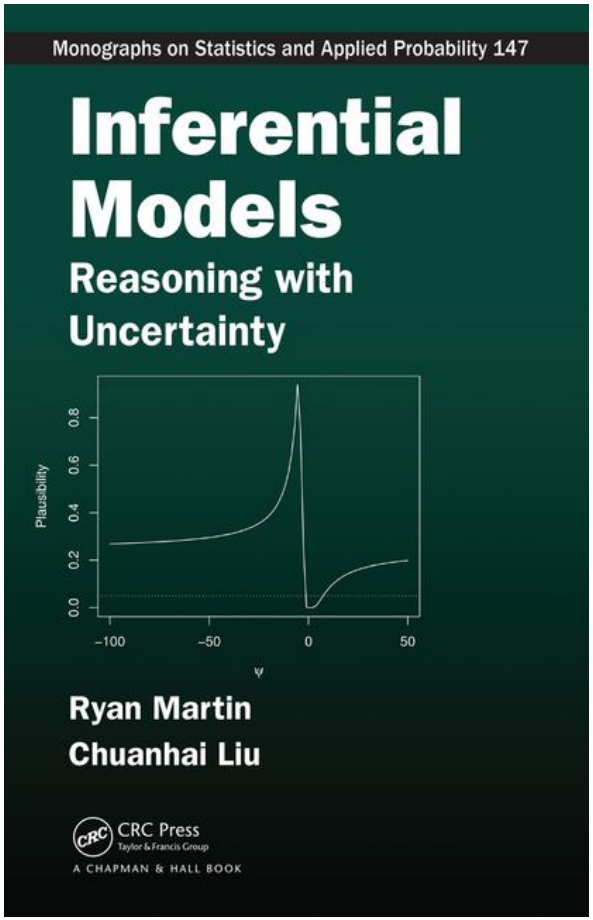
[2] All the legendary Grail quests involve tricks and riddles — don't take anything at face value, question/challenge everything!

Case 3, cont.

- Recognition that something other than the usual probability might be/is needed is the secret to finding the Grail.
- We have developed a new *inferential model* (IM) framework, built around the use of
  - random sets and
  - belief/plausibility functions.
- Some highlights:
  - provably valid/meaningful probabilistic inference
  - reduces to Bayes/probability when prior/model is given
  - *Vague Claim 2:* IM-based methods are never worse than existing methods.
- Check out references for details...

Case 3, cont.

Looking for that perfect holiday gift?



Makes a great stocking-stuffer!

Case 3, cont.

- A (light?) topical illustration I'm thinking about now.[3]
- Lots of talk recently about failure of election predictions.
- Simple setup: polling data X, outcome $Y \in \{\mathrm{C}, \mathrm{T}\}$.
- Predictions are based on probabilities

$$\pi_X(y), \quad y \in \{\mathrm{C}, \mathrm{T}\}.$$

- e.g., Nate Silver said $\pi_X(\mathrm{C}) = 0.72$ and $\pi_X(\mathrm{T}) = 0.28$.
- Are these probabilities meaningful...?

[3] With Prof. Harry Crane at Rutgers.

Case 3, cont.

- Let P be the joint distribution for (X, Y).
- Natural to ask that $x \mapsto \pi_x$ be *valid*, i.e.,

$$\mathsf{P}\{\pi_x(Y) \leq \alpha \mid X = x\} \leq \alpha, \quad \text{all } x, \quad \text{all } \alpha \in (0, 1).$$

- i.e., predictive probability assigned to the actual winner won't tend to be too small.
- If P is known, then the only reasonable choice is

$$\pi_x(y) = \mathsf{P}(Y = y \mid X = x).$$

- Easy calculation shows that validity holds:[4]

$$\begin{aligned}\mathsf{P}\{\pi_x(Y) \leq \alpha \mid X = x\} &= 1_{\pi_x(\mathrm{C}) \leq \alpha}\, \pi_x(\mathrm{C}) + 1_{\pi_x(\mathrm{T}) \leq \alpha}\, \pi_x(\mathrm{T}) \\ &\leq \alpha.\end{aligned}$$

[4] Recall *Vague Claim 1* above...

Case 3, cont.

- In reality, P is *unknown*, so the predictive probabilities $\pi_x(y)$ reported are based on model assumptions.
- Consequently:
  - Predictions don't reflect the uncertainty about P.
  - Validity argument fails — we can't use P to define $\pi_x(y)$.
- To account for model uncertainty, consider

$$\pi_x(y) = \sup_{\mathsf{P}} \mathsf{P}(Y = y \mid X = x),$$

  where "sup" is over all candidate models.
- Validity argument holds, but $\pi_x(y)$ is *not a probability!*

Case 3, cont.

- Way-too-simple example: Suppose we poll 1000 people and data is $x = (475$ for C, 425 for T, 100 non-response).
- Naive approach: ignore non-response and set

$$\pi_x(\mathrm{C}) = \tfrac{475}{900} = 52.8\% \quad \text{and} \quad \pi_x(\mathrm{T}) = \tfrac{425}{900} = 47.2\%.$$

- Ignoring non-response is a model assumption...
- To protect against uncertainty about non-response, try

$$\pi_x(\mathrm{C}) = \tfrac{475+100}{1000} = 57.5\% \quad \text{and} \quad \pi_x(\mathrm{T}) = \tfrac{425+100}{1000} = 52.5\%.$$

- These aren't probabilities!
  - $\pi_x(y) =$ "*plausibility* of y winning based on data x."
  - Based on available information, *both* candidates have high plausibility of winning — seems reasonable to me...

Case 3: Remarks

- Some important steps toward the Grail have been made.
- Lots of work left to do:
  - methods
  - computation
  - applications
  - theory and other fundamental developments.
- Quite literally, you can pick any area of statistics (spatial, survival, etc), work out an IM solution, and write a paper.[5]
- Other open questions:
  - Incorporating partial prior information?
  - Valid assessment of model uncertainty?
  - High-dim problems, penalties, etc?
  - Meta-analysis type things?

[5] Recall *Vague Claim 2* above...

Case 3: References

- M. and Liu (2013). Inferential models... *Journal of the American Statistical Association*.
- M. and Liu (2015). Conditional inferential models... *Journal of the Royal Statistical Society, Series B*.
- M. and Liu (2015). Marginal inferential models... *Journal of the American Statistical Association*.
- M. and Lingham (2016). Prior-free probabilistic prediction... *Technometrics*.
- M. (2016+). On an inferential model construction... *Journal of Statistical Planning and Inference*.
- M. and Liu (2016+). Validity and the foundations of statistical inference. `arXiv:1607.05051`.

The end

Thank you!