MUSINGS ON BAYESIAN NONPARAMETRICS

DAVID DUNSON, DEPARTMENTS OF STATISTICAL SCIENCE & MATHEMATICS

DUKE UNIVERSITY

DUNSON@DUKE.EDU

DISCLAIMERS

- This is not a tutorial on Bayesian nonparametrics
- I won't review basic Bayesian nonparametric modeling & computational approaches, such as Gaussian processes, Dirichlet process mixtures etc
- I've done that many times before & would find it boring & much of this audience (including students) has already been working in BNP
- Instead I'll try to step back & mull over what motivates nonparametric Bayes inference & whether BNP delivers on this promise & what the alternatives are

TUTORIAL OUTLINE

- Day 1: Bayes & Model Misspecification: Is “nonparametric” the answer?
- Day 2: Avoiding nonparametric (infinite-dimensional) models – some strategies
- Day 3: Adventures in clustering

DAY 1: BAYES & MODEL MISSPECIFICATION: IS "NONPARAMETRIC" THE ANSWER?

- I'll start with a *brief* review of parametric Bayes inference & what happens when the presumed model is wrong
- I'll discuss some strategies for dealing with model misspecification, including modeling averaging + lead into a motivation for nonparametric Bayes
- I'll then highlight what is good & bad about current nonparametric Bayes practice & motivate the need for new thinking in the field

PARAMETRIC BAYES – A CARTOON

- Bayes inference typically starts with choosing a likelihood function $L(y^{(n)}; \theta)$
- Let $y^{(n)}$ denote the observed data & θ parameters characterizing the likelihood of these data
- ***Parametric*** models are characterized by finitely-many θ parameters
- Parametric models are restrictive by definition – for example, requiring the density to have a very particular shape (e.g, Gaussian)
- However, with this restrictiveness comes (at least sometimes) simplicity in interpretation & computation

PARAMETRIC BAYES – THE PRIOR & POSTERIOR

- Once we have chosen a likelihood function $L\left(y^{(n)};\theta\right)$, we are ready to choose a prior for the parameters $\pi(\theta)$
- For interpretable parametric models, it becomes more feasible to choose a reasonable prior characterizing outside knowledge about the parameters (θ)
- We then update the prior with the likelihood function to obtain the posterior:

$$\pi\left(\theta|y^{(n)}\right) = \frac{\pi(\theta)L\left(y^{(n)};\theta\right)}{\int \pi(\theta)L\left(y^{(n)};\theta\right)d\theta} = \frac{\pi(\theta)L\left(y^{(n)};\theta\right)}{L\left(y^{(n)}\right)}$$

BAYESIAN INFERENCES

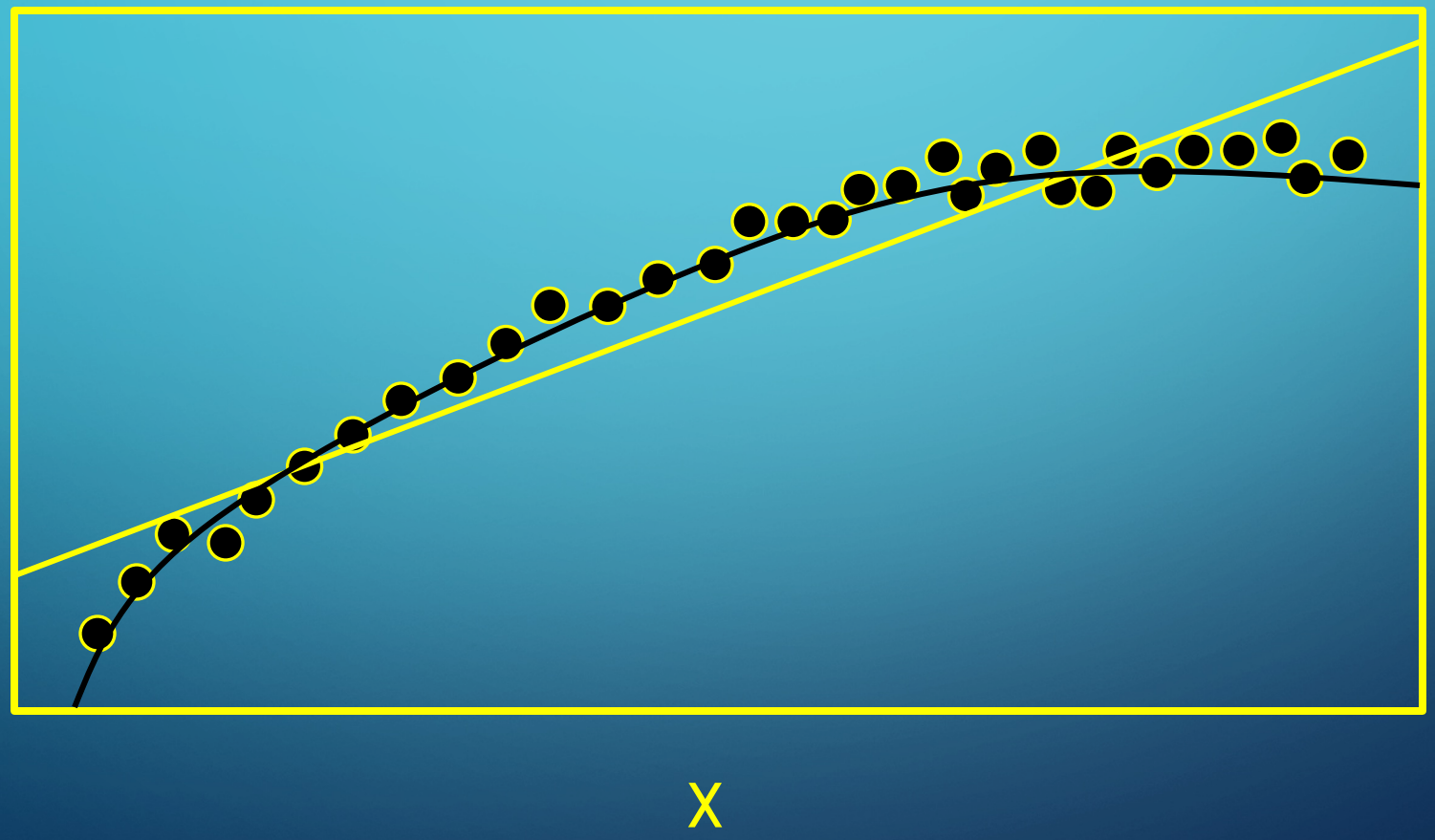
- With the posterior in hand (perhaps after substantial MCMC hurdles!), we can proceed to conduct Bayesian inferences
- This may involve presenting posterior summaries of the model parameters θ and functionals $g(\theta)$ of interest
- Or take a decision-theoretic approach, choosing a loss function $\ell(\delta, \theta)$ & solving $\hat{\delta} = \arg\min_{\delta} \ell(\delta), \ell(\delta) = \int \ell(\delta, \theta)\ \pi\left(\theta | y^{(n)}\right) d\theta$
- Again, for parametric models the above tends to be relatively straightforward

BUT WHAT IF THE MODEL IS WRONG??

- By the model being wrong, I mean that the *likelihood is misspecified*
- It is common knowledge that real world data never precisely follow any parametric model – hence, **all models are wrong**
- Often the inevitable imperfection of any parametric model is used as an argument in favor of nonparametric models
- However, wait a second – you loose a lot in abandoning a parametric model (which is relatively easily interpretable and computable) in favor of an (often black box and immensely complicated) nonparametric alternative!

A TOY EXAMPLE

- Comparing a nonparametric curve fit to a linear model fit



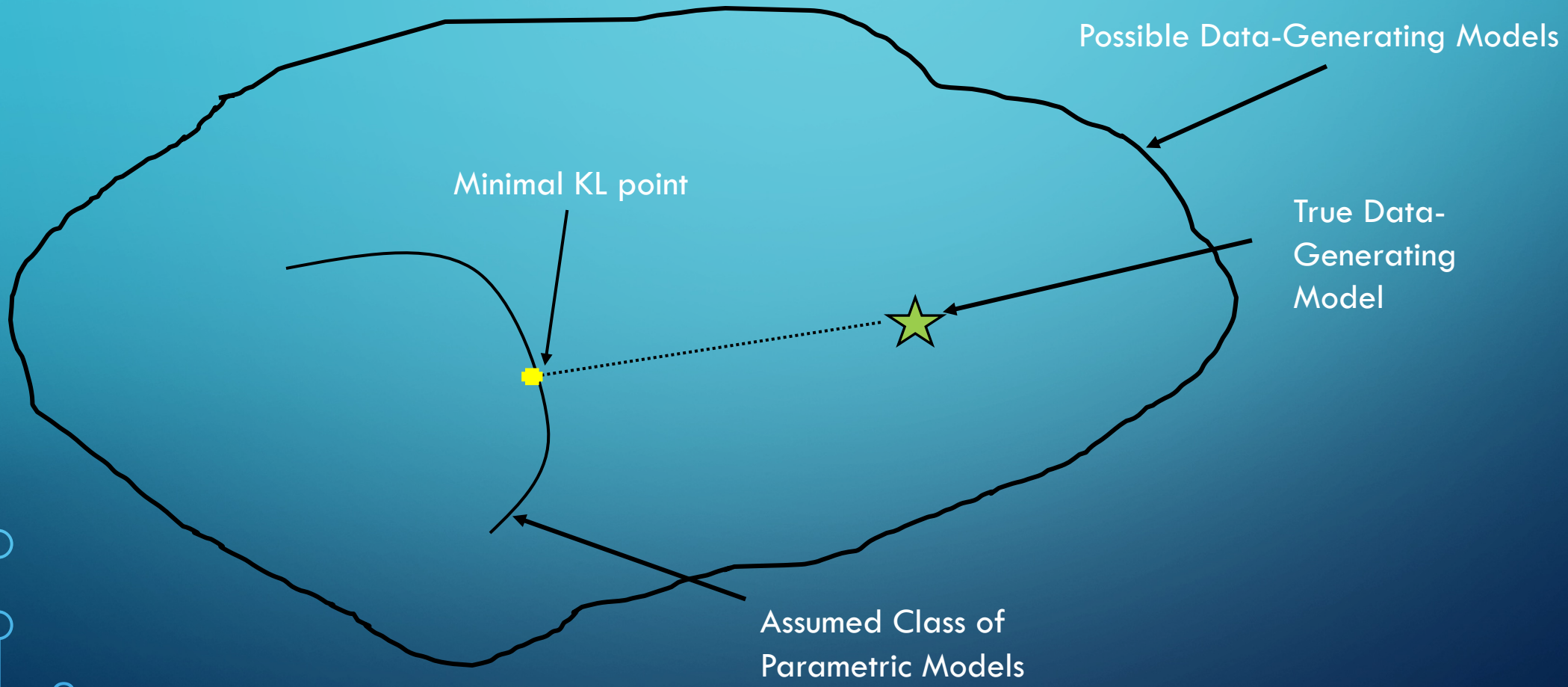
COMMENTS ON TOY EXAMPLE

- It seems clear that the data weren't generated from a linear model
- However, we knew that in advance under the all models are wrong credo
- The linear model does provide a decent "coarse" approximation to the curve shape
- The linear model is also massively easier to summarize as we just have an intercept and slope
- In 1d the nonparametric fit can be visualized but as dimension increases that becomes impossible – how to provide an interpretable summary?

WHY NOT JUST USE THE PARAMETRIC MODEL?

- It is common to ignore some amount of misspecification in the chosen parametric model as long as it is not “too large”
- How does the posterior distribution under our parametric model behave if the model is wrong?
- Can we trust our point and interval estimates of parameters & our predictions from a flawed parametric model?
- There is some associated theoretical understanding that may help address these questions

A CARTOON OF MISSPECIFICATION



BEHAVIOR OF POSTERIOR UNDER MISSPECIFICATION

- One can characterize how close $L(y^{(n)};\theta)$ is to the true data-generating model $f_0(y^{(n)})$ using Kullback-Leibler (KL) divergence
- The parametric class is a measure zero subset of the set of possible data-generating likelihoods (*represented by the curve in the cartoon*)
- Let θ^* denote the value of θ at which the KL divergence from $f_0(y^{(n)})$ to $L(y^{(n)};\theta)$ is minimized – this is the **pseudo-true** parameter value

BERNSTEIN VON MISES & MISSPECIFICATION

- Under some regularity conditions, the posterior $\pi(\theta|y^{(n)})$ tends to converge as sample size increases to a multivariate Gaussian distribution
- Such results are referred to as Bernstein von Mises theorems
- Show an asymptotic equivalence between frequentist maximum likelihood-based inferences and Bayes
- In the misspecified case, the posterior concentrates around the pseudo-true parameter value θ^*

TROUBLES ARISE

- Initially this appears to be a good result – if the model is misspecified then we just target the pseudo-true parameter value instead of the true value
- The posterior under the presumed parametric model will tend to concentrate around this value so all is good?
- However, if the model is misspecified, the asymptotic variance is wrong – motivating so-called sandwich adjustments to frequentist ML-based estimators
- Also, there is the question of interpretation – if the model is wrong, can we reliably interpret the estimated parameters as if the model were true?

IS KL A GOOD CHOICE?

- Also targeting the pseudo-true parameter value presumes the KL is a good choice of loss
- KL tends to be very sensitive to misspecification in the tails of the distribution
- For example if the model misspecification corresponds to a small proportion of the sample being corrupted, the minimal KL point may be sensitive to these outliers
- Also, there may be multiple minimal KL points or one may be able to minimize the KL in a parametric class by increasing the number of parameters

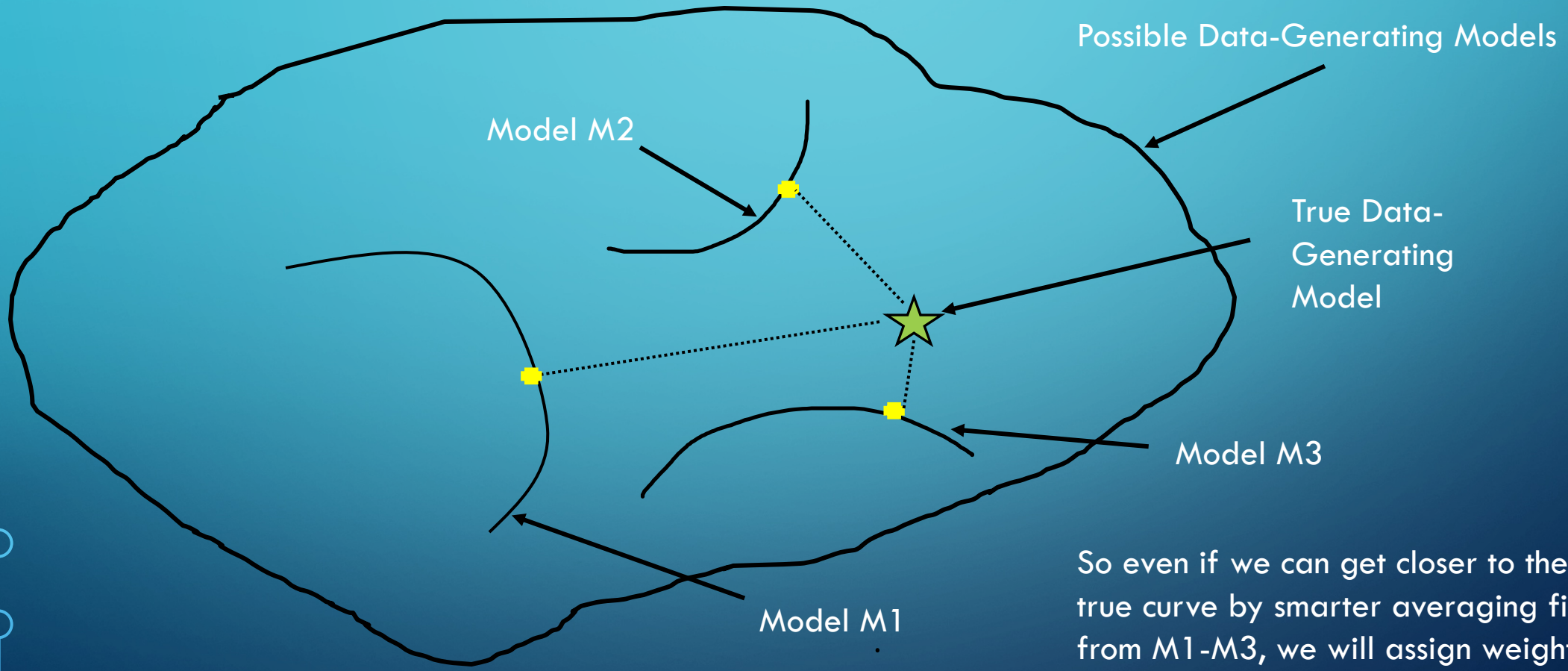
SO WHAT TO DO?

- Before abandoning parametric model inference, how about trying Bayesian model averaging and/or selection?
- So instead of focusing on a single model, we consider a list of possible models $\mathcal{M} = \{\mathrm{M}_1, \ldots, \mathrm{M}_k\}$ & define a posterior over these k models
- To do this we need a prior $\pi(\mathcal{M})$ placing probabilities on each model, as well as priors for the coefficients θ_j in model M_j for $j = 1, \ldots, k$
- Then we can define posterior model probabilities: $pr(\mathrm{M}_j|y^{(n)}) = \frac{pr(\mathrm{M}_j)L(y^{(n)}|\mathrm{M}_j)}{\sum_h pr(\mathrm{M}_h)L(y^{(n)}|\mathrm{M}_h)}$

BAYESIAN MODEL AVERAGING IN TOY EXAMPLE

- In addition to the linear model M_1, we may include a logistic growth curve model M_2 and a few other alternatives $\mathrm{M}_3, \mathrm{M}_4, \mathrm{M}_5$ in our set of models $\mathcal{M}$
- By optimally averaging these simple models, we can obtain a good approximation to the true non-linear curve
- However, the weights in Bayesian model averaging are $pr\left(\mathrm{M}_j \middle| y^{(n)}\right)$ - the posterior probability that M_j is the true data-generating model f_0!
- As sample size increases, this weight $\rightarrow 1$ for the model that is closest in KL to the true model f_0
- This is due to the flawed M-closed assumption that f_0 is one of the k models in $\mathcal{M}$

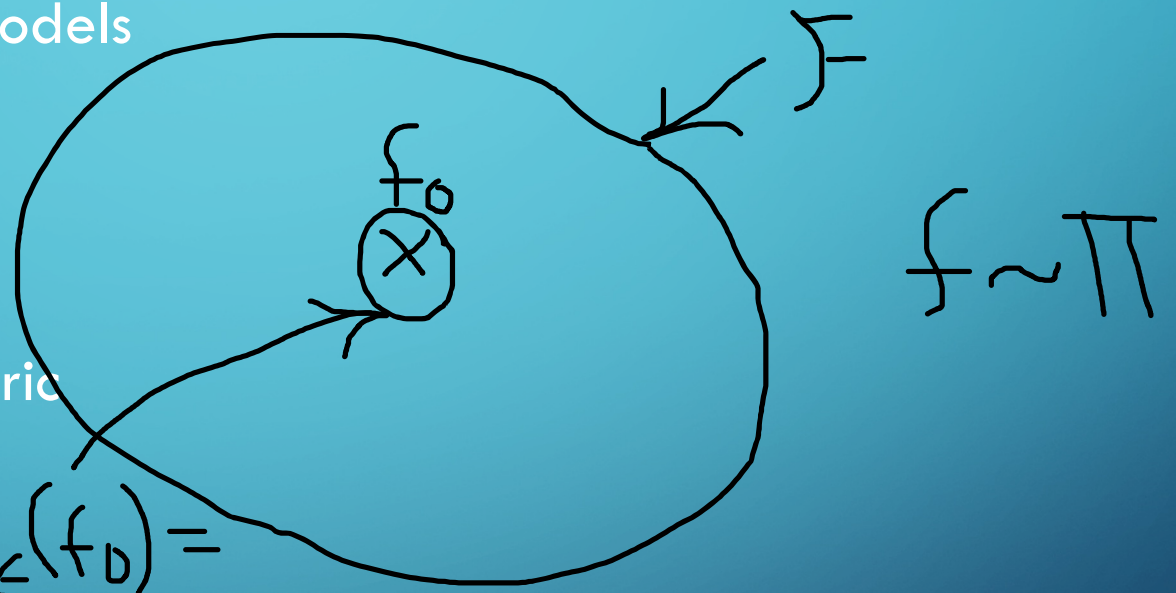
A CARTOON OF MISSPECIFICATION – MA CASE



So even if we can get closer to the true curve by smarter averaging fits from M1-M3, we will assign weight 1 on M3 as n increase

BAYESIAN NONPARAMETRIC PHILOSOPHY

- $\mathcal{F}$ = "big" set of possible models
- f_0 = unknown true model
- $f \sim \Pi$ = prior for model
- Large support=nonparametric



$\Pi\{\eta_\epsilon(f_0)\} > 0$

$\forall \epsilon > 0,$

$f_0 \in F$

$\eta_\epsilon(f_0) =$

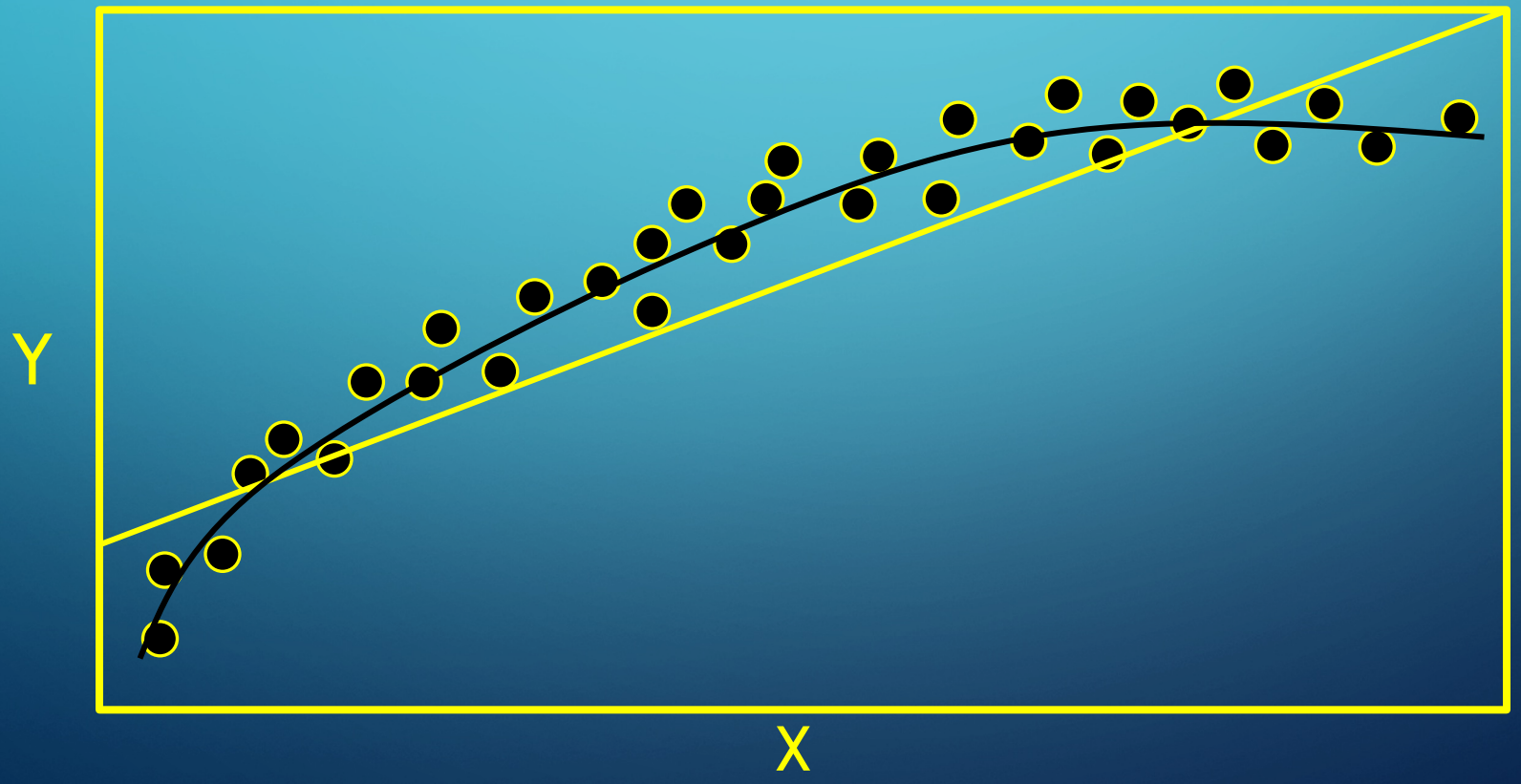
$\{f : d(f_0, f) < \epsilon\}$

NONPARAMETRIC BAYES = LARGE SUPPORT

- So to be "truly" nonparametric in a Bayesian sense, we need a prior Π for an infinite-dimensional unknown, such as a regression surface or a density
- This prior needs to be able to generate $f \sim \Pi$ in such a manner that there is positive probability of being arbitrarily close to any f_0 in a broad class $\mathcal{F}$
- Also, we would want Π to be as simple as possible in terms of prior elicitation and posterior computation – otherwise, applications may be intractable
- It is of course hard to come up with appropriate choices of Π leading to a lot of focus in the literature on similar approaches (GPs, DPMs, etc)

A TOY EXAMPLE - REVISITED

- $y_i = f(x_i) + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2), i = 1, \ldots, n$
- $f \sim GP(m, c)$ = Gaussian process prior w/ mean m & covariance c



HOW TO INTERPRETE THE RESULTS?

- In the one dimensional nonparametric case of the toy example, we can present point & interval estimates for the curve
- For example, the posterior mean $\hat{f}(x)$ and pointwise (or simultaneous) 95% credible bands
- If $x = (x_1, \ldots, x_p) \in \mathbb{R}^p$, with $p \geq 3$, then we can't visualize the estimated regression function not to mention the posterior for this function
- If we want a black-box predictive model than we don't care but BNP is perhaps not the best pragmatic choice for black box prediction

BACKING OUT INTERPRETABILITY

- There is a recent literature attempting to back out interpretability in a **decision theoretic stage** after obtaining $\Pi(f|y^{(n)})$
- Woody et al (2021) propose to fit simple interpretable models to the posterior samples of $f(\mathrm{x})$
- Kowal et al (2021) use a related philosophy for subset selection and inference on variable importance
- Afrabandpey et al (2020) fit a black-box Bayesian predictive model in a first stage & then select an interpretable model that is close to the black box

SOME COMMENTS ON LECTURE 1

- First fitting a Bayesian nonparametric model & then attempting to back out interpretability in a decision theoretic phase is more principled
- We avoid assuming that an overly simple model provides a realistic characterization of the true data-generating model
- Instead we view finding a simple surrogate model as a decision problem
- We can also view finding simple summaries as decision problems
- In general, Bayesians too often ignore the decision analysis phase!

DISCUSSION

- While we need new BNP models & computational algorithms – e.g, for high-dimensional data, new types of data etc – there is perhaps a more pressing need for improving interpretability, ease of implementation & reproducibility
- There is a lack of code for routine implementation, many BNP algorithms are "flaky" (MCMC may fail etc), adaptations are needed for complexities that arise for real data (missingness, etc) & interpretability is problematic
- For these reasons, BNP methods are not used routinely in practice with some notable exceptions – BART, GP for spatial random effects, etc

TAKE HOME MESSAGES – LECTURE 1

- We need new general-use (actually reliable & efficient!) software that produces interpretable results
- BART has been a dramatic success due to the very efficient & reliable code – even though it's a black box for prediction, the code motivates a lot of focus
- In contrast, how to efficiently implement density estimation & associated inferences? There are tons of methods but a real lack of useful code
- To be relevant in the broader scientific community & move beyond a small niche area, we need to do this!

TAKE HOME MESSAGES – CONTINUED

- BNP methods have a reputation of being flaky & non-interpretable & slow
- We need to address these problems & improve issues of brittleness & lack of reproducibility & scalability
- New approaches for obtaining interpretable results are needed
- One big issue is the need to model everything – this is inherent in the Bayesian framework & a BNP model of everything is daunting to specify & compute!
- Can we think of new ways to acknowledge model misspecification without modeling everything with a giant complicated infinitely-parametric model?

LECTURE 2: AVOIDING NONPARAMETRIC BAYES

- Usual nonparametric Bayes involves specifying a likelihood involving an infinite-dimensional parameter to induce sufficient flexibility
- Hence, it is common to refer to “nonparametric Bayes” as a misnomer
- It is very different from common frequentist nonparametric methods that do not ”model everything” but reduce focus to ranks etc
- Can we take a Bayesian approach to inference that acknowledges model misspecification but that doesn’t model everything?

AVOIDING BNP – SOME STRATEGIES

- A common strategy is to apply a generalized Bayes (G-Bayes) approach, which replaces the likelihood function with some alternative
- Pseudo-likelihood, rank likelihoods, composite likelihoods etc etc
- My favorite example: Hoff (2007) "Extending the rank likelihood for semiparametric copula estimation" AOAS 1(1), 265-283.
  - This is a super useful approach that avoids specifying a full likelihood for the marginals
  - A Gaussian copula model is used to characterize dependence
  - A simple Gibbs sampler can be used for computation & an R package is available

SOME OTHER EXAMPLES OF BAYES WITH BOGUS LIKELIHOODS

- Yu & Moyeed (2001) Bayesian quantile regression – asymmetric Laplace "likelihood" is chosen to mimic loss function for quantile regression
- Duan & Dunson (2021) use a purposely overly-simplified "spanning tree likelihood" to robustly infer the backbone of the dependence graph
- Dunson & Taylor (2005) proposed a substitution likelihood for quantiles, which avoids modeling the density of the data between prespecified quantiles
- There are also many papers on Bayesian implementations of composite likelihoods, partial likelihoods, pseudo likelihood, etc

QUESTIONS ABOUT BAYES WITH BOGUS LIKELIHOODS

- Often such approaches have good practical performance in simulations & real data applications
- Typically, theoretical justification relies on an appeal to frequentist asymptotics, showing consistency, convergence rates, BvM, etc
- Hard to answer the question about the interpretation of the resulting (not quite) posterior – what do uncertainty statements mean etc?
- Hard to justify fully from a Bayesian perspective – though one could argue that typical Bayes justifications for parametric & nonparametric models are also flawed for different reasons

COHERENT BAYES INFERENCE WITHOUT LIKELIHOODS

- So is there any hope of defining a coherent Bayesian inferential framework without a full likelihood specification of "everything" (all aspects of the data)?
- An exciting development in this regard is the class of Gibbs posteriors
- In defining a Gibbs posterior, we replace the log-likelihood $\log L\left(y^{(n)};\theta\right)$ with a loss function $\ell\left(\theta;y^{(n)}\right)$:

$$\pi_{\psi}\left(\theta|y^{(n)}\right)=\frac{\pi(\theta)e^{-\psi\ell\left(\theta;y^{(n)}\right)}}{\int\pi(\theta)e^{-\psi\ell\left(\theta;y^{(n)}\right)}d\theta}$$

GIBBS POSTERIORS

- $\pi_\psi\left(\theta|y^{(n)}\right) =$ is then treated essentially as a usual posterior for the parameter θ in conducting Bayesian inference
- There is no generative model here as we bypass a likelihood specification entirely!
- There is great freedom in choosing a rich variety of loss functions depending on the applied context & essentially independent from data modeling
- The scalar tuning parameter ψ plays a key role in controlling concentration of the Gibbs posterior & hence posterior uncertainty (eg., Martin & Syring 2022)

GIBBS POSTERIOR – DECISION THEORETIC JUSTIFICATION

- But isn't this all sort of ad hoc – similarly to using composite or pseudo likelihoods in place of a "real" likelihood in Bayes rule?
- Actually there is a beautiful paper Bissiri et al (2016) providing a formal Bayesian decision theoretic justification for Gibbs posteriors
- The target of inference for the Gibbs posterior is not some "true" parameter value in a likelihood function, but is

$$\theta_{opt} = \arg\min_{\theta} \mathbb{E}_{\pi_0}\{\ell(\theta; y^{(n)})\}$$

π_0 = true data-generating process, above expectation = frequentist risk

GIBBS POSTERIOR – CONTINUED

- The Gibbs posterior is the "best" conditional distribution for quantifying our subjective beliefs about θ_{opt} in the absence of a likelihood function
- There's no doubt this is a useful framework, with the main practical questions being how to choose the loss & (particularly) the tuning parameter ψ
- For a recent example, using Gibbs posteriors for clustering including for uncertainty quantification of k-means, refer to Rigon et al 2020
- BUT throwing out the likelihood entirely & focusing completely on loss-based inferences sacrifices many of the advantages of Bayesian inference!

HOW TO KEEP THE "MODEL" IN BAYESIAN INFERENCE WITHOUT NEEDING TO BELIEVE ITS EXACTLY TRUE?

- A major problem with Bayesian inference is the implicit belief that the presumed model is exactly true – this can leads to pitfalls/brittleness
- There are a number of approaches available for relaxing this assumption
- Approach 1: pragmatically be aware that your model may be (at least) slightly wrong but hope that this misspecification doesn't lead to practical problems; indeed the behavior I discussed in lecture 1 is reassuring (for many cases)
- For example, suppose we model the data as $y_i \sim N(\mu, \sigma^2)$ but the true distribution is slightly non-Gaussian. Then, inferences on the mean & variance will tend to be robust

PRAGMATIC BAYES – CONTINUED

- In many other Bayesian models, we will also be robust to slightly misspecified likelihoods
- In such cases, a practical Bayesian may consider a variety of models, conduct goodness-of-fit assessments & choose the model having the "best" balance of parsimony, interpretability & fit to the data
- The hope is then that the final selected model will be only slightly misspecified and inferences will be robust to that misspecification

BRITTLENESS OF PRAGMATIC BAYES

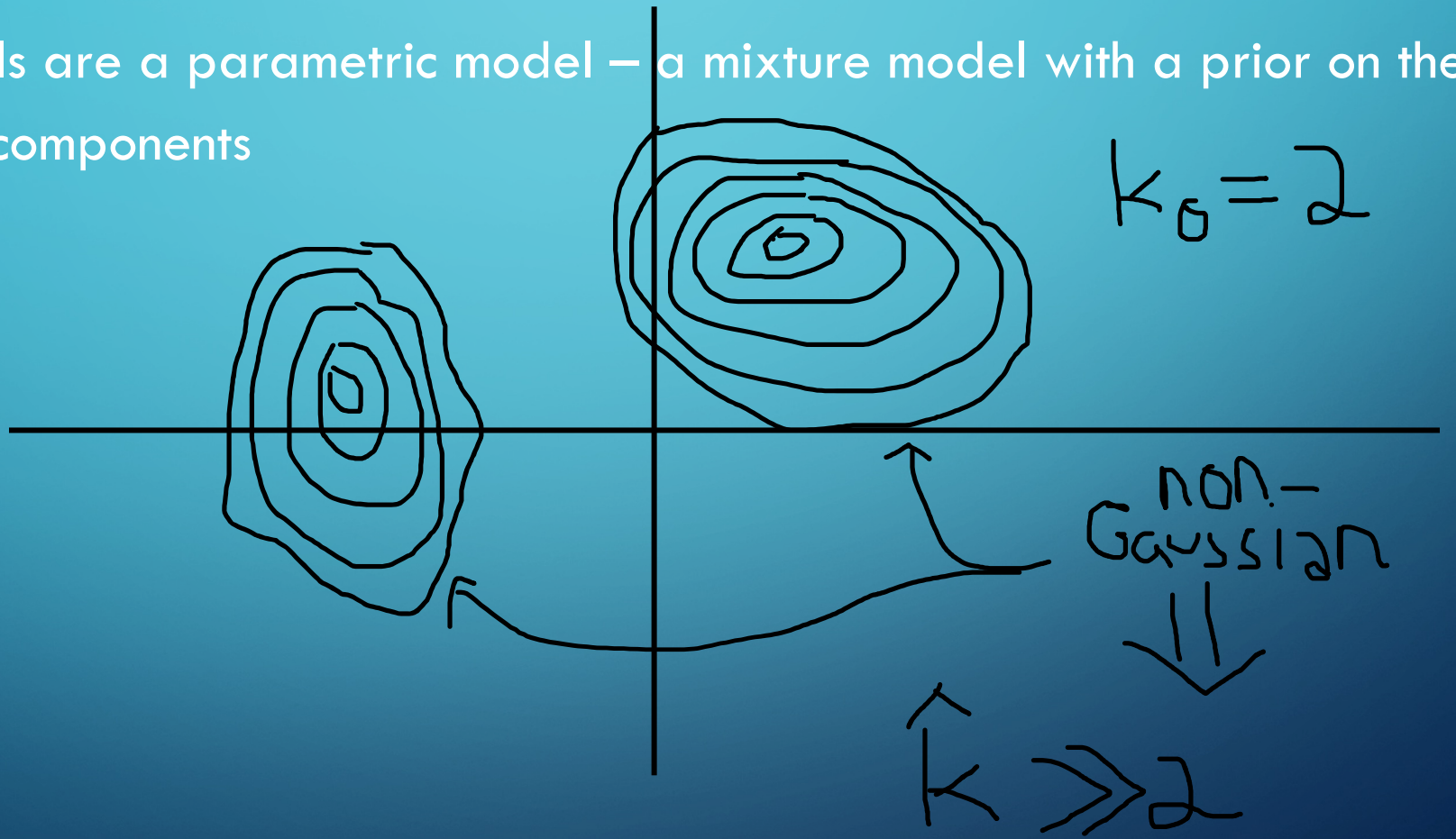
- Although many types of inferences do tend to be robust to slight model misspecification, there are certain type of inferences which are not
- One such class includes cases in which there is a model complexity index k, which is either formally treated as unknown through a prior or otherwise chosen based on the data.
- In many such cases, even slight misspecification can lead to considerable over-estimation of k, with $\hat{k} \to \infty$ as $n \to \infty$

BRITTLENESS EXAMPLE - CLUSTERING

- The typical Bayesian solution to clustering problems relies on finite mixture models: $f(y) = \sum_{h=1}^{k} \pi_h \mathcal{K}(y; \theta_h)$, where k is the number of clusters, π_h is the weight on cluster h, and $\mathcal{K}(\theta_h)$ describes variability within cluster h
- There is a huge literature choosing all sorts of clever priors for $\{\pi_h\}$ and an active debate on infinite mixtures vs finite mixtures & inconsistency questions
- Miller and Harrison (2018) show inconsistency of Dirichlet process mixtures (DPMs) for selecting the number of clusters, while advocating for the use of mixture of finite mixtures (MFMs)

MIXTURES OF FINITE MIXTURES & BRITTLENESS

- MFM models are a parametric model – a mixture model with a prior on the number of components



EVEN SMALL MODEL MISSPECIFICATION HAS BIG CONSEQUENCES

- Slight misspecification of the Gaussian kernel assumptions leads a ”true” clusters to be broken up into smaller artifactual sub-clusters
- This demonstrates a type of brittleness to model misspecification
- Such brittleness often occurs when complexity of the model is allowed to be adaptive to data – unknown lag in time series, # of predictors in regression, # of basis functions, # of factors, model comparison for nested models, etc

HOW TO REDUCE BAYESIAN BRITTLENESS?

- A number of ideas have been proposed to reduce brittleness of Bayesian inferences
- As we motivated earlier, in the misspecified case, Bayesian inference targets the pseudo-true parameter minimizing the KL divergence from the true model
- The choice of KL can contribute to the brittleness problems
- Jewsen et al (2018) propose to replace the KL with alternative divergences to improvement robustness to model misspecification
- Very cool & rich idea, but can lead to computational challenges

REDUCING BAYESIAN BRITTLENESS – CONT …

- An alternative idea is C-Bayes (Miller & Dunson, 2019), which is designed to define an alternative "Coarsened" posterior which explicitly acknowledges a small amount of model misspecification
- The standard posterior conditions on the event that the observed data are generated by sampling from the model – incorrect under misspecification
- Rough idea: condition on the event that the empirical distribution of the observed data is close to the empirical distribution of data sampled from the model

C-POSTERIOR – SETTING THE STAGE

- Suppose we have a model $\{P_\theta, \theta \in \Theta\}$ & a prior Π on Θ
- Let $\theta_I \in \Theta$ represent the *idealized distribution* of the data
- Suppose $X_{[1:n]}$ are drawn iid from P_{θ_I} but we don't observed these data
- Observed data $x_{[1:n]}$ are slightly corrupted version of $X_{[1:n]}$ in the sense that $d\left(\hat{P}_{X_{[1:n]}}, \hat{P}_{x_{[1:n]}}\right) < r$, for $\hat{P}_{x_{[1:n]}} = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}$, distance $d(.,.)$ & $r > 0$.
- $x_{[1:n]}$ behave like iid samples from P_0 but due to corruption $P_0 \neq P_{\theta_I}$

C-POSTERIOR – DEFINITIONS

- If there were no "corruption" & the model was correct, then we should use the standard posterior & condition on the event that $X_{[1:n]} = x_{[1:n]}$
- Alternatively, one could attempt to model the "corruption" – but this just leads to another more complicated model which is also inevitably misspecified
- Rather than the standard posterior $\Pi(d\theta | X_{[1:n]} = x_{[1:n]})$ we use c-posterior:

$$\Pi\big(d\theta | d_n\big(X_{[1:n]}, x_{[1:n]}\big) < R\big),$$

Where $R \sim H$ is a prior on the radius & $d_n\big(X_{[1:n]}, x_{[1:n]}\big) \geq 0$

INTUITION & RELATIVE ENTROPY C-BAYES

- The idea here is that we are formally allowing a “small” amount of model misspecification and/or corruption in the data
- This will hopefully allow us to use a parametric model without the brittleness & other issues
- The type of robustness depends on the choice of distance/discrepancy
- For practical reasons Miller & Dunson (2019) focus on relative entropy,

$$d(P_\theta, P_0) = \int p_0(x) log \frac{p_0(x)}{p_\theta(x)} \lambda(dx)$$

PRACTICAL FORM OF C-POSTERIOR

- Based on some theoretical arguments found in Miller & Dunson (2019),

$$\Pi\left(d\theta \middle| d_n\left(X_{[1:n]}, x_{[1:n]}\right) < R\right) \propto \Pi(d\theta)\{\prod_{i=1}^{n} p_\theta(x_i)\}^{\zeta_n},\ \zeta_n = \frac{1}{1+n/\alpha},$$

- Here, we are assuming $R \sim Exp(\alpha)$ so that α controls the neighborhood size
- A key advantage relative to related work & alternative choices of distance, is that we bypass dependence of the posterior on a density estimator
- Interestingly, the c-posterior involves a power likelihood – this tends to make Bayes computation easy & relates to a rich literature on power likelihoods

SOME NOTES ON C-BAYES

- For large samples n the standard posterior can be strongly affected by small changes to the observed data distribution P_0, particularly for model inference
- In contrast, c-posteriors are robust to small changes in P_0
- Although we focused on iid, time series, regression applications are direct
- Examples in paper – kth order auto-regression with unknown k, variable selection in linear regression, mixture models with prior on # components
- C-Bayes provides a simple way to robustify parametric model inferences

OTHER WAYS TO ROBUSTIFY PARAMETRIC BAYES

- We have seen that "model inferences" in which we have some complexity index k present some of the biggest issues in terms of brittleness
- I'm ruling out modeling the parametric model misspecification, as that just leads to another model
- One possibility is to go from M-closed (model is exactly right) & M-open (model is wrong) to M-complete (inferential models are wrong but we have a correct reference model)
- Li and Dunson (2020) proposed such an approach for model selection & averaging of linear models using Gaussian process (GP) regression as the reference

MODEL AVERAGING

- Usual posterior model probabilities: $pr(\mathrm{M}_j|y^{(n)}) = \frac{pr(\mathrm{M}_j)L(y^{(n)}|\mathrm{M}_j)}{\sum_h pr(\mathrm{M}_h)L(y^{(n)}|\mathrm{M}_h)}$
- $pr(\mathrm{M}_j|y^{(n)})$ = posterior probability the model is exactly correct
- There is an existing literature on Bayes model selection from M-open or M-complete taking decision theoretic perspective & using cross validation (e.g, Clyde & Iversen, 2013; Gutierrez-Pena et al 2009).
- Alternatively, one can do non-Bayesian aggregation of models – e.g., Rigollet & Tsybakov (2012) develop an exponential weighting approach

ABSOLUTE & RELATIVE MODEL WEIGHTS

- Let $\mathcal{N}^*$ be an oracle model that generated the data with f^* the density
- For model M_j with density f_j define absolute model weights $\pi_j = e^{-nKL(f^*, f_j)}$
- π_j = probability of selecting M_j under a randomized decision rule in the absence of sufficient evidence in the data to distinguish M_j from $\mathcal{N}^*$
- We can also define conditional model weights by normalizing π_j over $\mathcal{M}$
- Weights are asymptotically equivalent to usual posterior model probabilities if one of the models in the list is true, so the oracle is one of the candidates

D-PROBABILITIES

- Of course in practice, we don't know the oracle – however, we can use a nonparametric Bayes model as a surrogate for the oracle
- We refer to the resulting model weights as D-probabilities, as they provide a divergence-based alternative to usual Bayesian posterior model probabilities
- Although the framework is general, the details work out particularly nicely when the models in $\mathcal{M}$ are linear & $\mathcal{N}$ is a Gaussian process (GP) regression
- In comparing models $t \in \{1,2\}$, the D-probabilities factorize as $\pi_{j,t} = e^{-\mathcal{G}_{j,t}-\mathcal{P}_{j,t}}$, where $\mathcal{G}_{j,t}$ is a goodness-of-fit term & $\mathcal{P}_{j,t}$ is a penalty on complexity

D-PROBABILITIES – CONTINUED

- $-\mathcal{P}_{j,t} = \frac{1}{2}(p_j + 1)$ or $-\mathcal{P}_{j,t} = \frac{\log 2}{2}(p_j + 1)$ with p_j = # predictors
- When comparing models, the penalty is identical to certain existing methods like AIC, the pseudo-Bayes factor & the posterior Bayes factors
- A closed form is obtained for the absolute & relative model probabilities up to a couple of covariance parameters, which can be estimated via EB/MCMC
- D-probabilities provide a strength of evidence for lack of fit

DISCUSSION ON "AVOIDING BNP" (LECTURE 2)

- It is important to acknowledge that any parametric model is not exactly true
- Particularly in model selection contexts (including for "encompassing" models including a complexity index k), even small model misspecification can have drastic impact
- This can be due to the minimal KL point in the parametric class having $k \to \infty$
- Not clear that Bayesian nonparametric models are the best solution
- There are many G-Bayes & other more robust approaches, which acknowledge model misspecification under M-open or M-complete

LECTURE 3: ADVENTURES IN CLUSTERING

- One of the canonical tasks of Bayesian nonparametric (BNP) models, and closely-related finite mixture models, is clustering
- BNP (in general) tends to be robust & have good performance as a flexible black box for density estimation, regression & prediction
- However, clustering tends to be a *much* more delicate exercise that depends critically on modeling assumptions – hence, BNP clustering methods are effectively parametric & share the brittleness etc issues to misspecification
- In this ”tutorial”, I describe some challenges, pitfalls & tentative (partial) solutions to practical issues that arise

BASIC MODEL-BASED CLUSTERING

- Model-based clustering typically relies on finite mixture models: $f(y) = \sum_{h=1}^{k} \pi_h \mathcal{K}(y; \theta_h)$, where k = # clusters, π_h = weight on cluster h & $\mathcal{K}(\theta_h)$ describes variability within cluster h
- This implies $y_i \sim \mathcal{K}(\theta_{c_i}), pr(c_i = h) = \pi_h$, where $c_i \in \{1, \ldots, k\}$=cluster id
- Mixtures of finite mixtures (MFMs) choose a prior on k along with the other unknowns – the component probabilities & cluster-specific parameters
- BNP approaches let $f(y) = \int \mathcal{K}(y; \theta\,)\, dP(\theta)$, where $P \sim \Pi$, P is an almost surely discrete random probability measure & Π is an appropriate prior

PRIORS FOR THE DISCRETE MIXING MEASURE

- Random probability measure P associated with discrete prior Π can be represented as $P = \sum_{j=1}^{\infty} \pi_j \delta_{\theta_j^*}$, where δ_θ is a degenerate distribution at θ & $(\pi_j)_{j \geq 1}$ is a sequence of non-negative r.v.s $\sum_{j=1}^{\infty} \pi_j = 1 \ as.$
- Atoms θ_j^* are sampled iid from diffuse base measure P^* independently of π_j
- This induces the infinite mixture model: $f(y) = \sum_{j=1}^{\infty} \pi_j \mathcal{K}(y; \theta_j^*)$.
- Clustering is induced from ties in the samples P. If we let $\theta_i \sim P, P \sim \Pi$, then with the above specification, we will get $\theta_i = \theta_{i'}$ with positive probability

CLUSTERING/PARTITIONING PRIORS

- Marginalizing over the prior Π induces a prior for the partition of n subjects into k groups.
- Assuming exchangeability, the prior probability on a particular partition of $[n]$ into k groups of sizes $n_1, \ldots, n_k$ only depends on k & $n_1, \ldots, n_k$
- This prior is represented via what is known as the exchangeable partition probability function (EPPF): $p_k^{(n)}(n_1, \ldots, n_k)$
- The specific form of EPPF depends on Π, with different choices favoring different behaviors in terms of growth in the number of clusters with n

GIBBS-TYPE – A BROAD FAMILY OF PRIORS

- Gibbs-type priors consist of species sampling models such that the EPPF is $p_k^{(n)}(n_1, \dots, n_k) = V_{n,k} \prod_{i=1}^{k} (1-\sigma)_{n_i-1}$, $\sigma < 1$
- $V_{n,k}$ are non-negative weights satisfying a simple recursion
- Special cases include DP, mixtures of DPs, Pitman-Yor etc (De Blasi et al 2015)
- Under Gibbs-type priors we obtain the following prediction rule:

$$P(X_{n+1} = \text{“}new\text{”} | X_1, \dots, X_n) = \frac{V_{n+1,k+1}}{V_{n,k}} = f(n, k, \Theta)$$

- For $\sigma < 0$ finite # clusters, $\sigma = 0$ $\log n$ growth, $\sigma \in (0,1)$ n^{σ} polynomial

SOME COMMENTS ON PRIOR CHOICE

- If we directly observe the distinct types (species) in a sample, then these processes provide a useful framework for modeling & prediction
- For a recent example, refer to Zito et al (2022) – showing also that the data can inform about differences between different species sampling priors
- If the distinct types are latent cluster ids, then both the form of EPPF & the prior P^* have an important impact on the clustering posterior
- However, it is less clear the extent to which the data can inform about these choices – practitioners often focus just on DPMs & MFMs favoring few clusters

CHOICE OF KERNEL

- The other component of the specification is the choice of kernel $\mathcal{K}(y;\theta)$
- The kernel describes the exact parametric distribution characterizing within-cluster variability
- There has been an active debate in the literature discussing clustering consistency when the true model is $f_0(y) = \sum_{h=1}^{k_0} \pi_{0h}\mathcal{K}(y;\theta_{0h}),$
- In general, this work focuses on the case in which the kernel is assumed to be correctly specified
- If the kernel is misspecified, then most models will break a true cluster into more & more clusters without bound as sample size increases

BAYESIAN "NONPARAMETRIC" CLUSTERING?

- It seems that in terms of the clustering structure, BNP priors are very much parametric and restrictive
- It also seems that these models (with a few notable exceptions in the literature) assume a parametric distribution for within cluster variability
- Hence, from the standpoint of clustering BNP models are effectively parametric models
- They induce flexible/robust models for the density that have nice asymptotic properties etc but the estimates of clustering are highly brittle

SO WHAT TO DO ABOUT THIS PROBLEM?

- One can try to model the within-cluster densities more flexibly, e.g., using nonparametric estimators with unimodality (e.g. Rodriguez & Walker 2014)
- However, defining kernels that are too flexible leads to ambiguity in defining clusters & identifiability issues (see Hennig et al 2015)
- Alternatively, we can fall back on some of the solutions we discussed for robustifying inference under parametric models – C-Bayes, Gibbs posteriors
- In the remainder, I'll illustrate some possibilities briefly

BAYESIAN DISTANCE CLUSTERING

- Most of the literature on clustering is based on defining pairwise distances between data points, and a corresponding loss function for clustering.
- This avoids the need to specify a full generative probability model for the data but is sensitive to the distance/loss & typically lacks UQ
- Duan & Dunson (2021) propose to define a Bayesian clustering model for a pairwise distance matrix instead of directly for the data
- The goal is to robustify model-based clustering without needing a complex model for the density of the data within each cluster

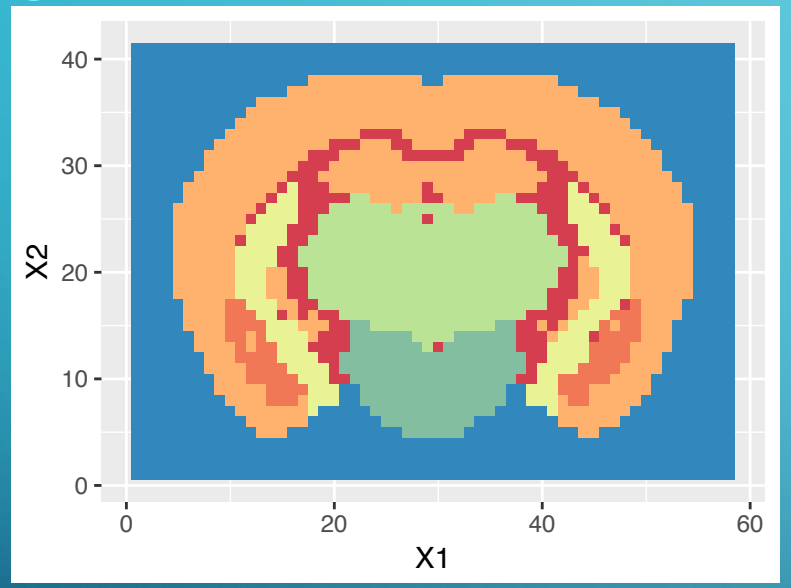
BAYESIAN DISTANCE CLUSTERING – CONTINUED

- Let $y_1^{[h]}$ = cluster seed, $\tilde{d}_{i,1}^{[h]} = y_i^{[h]} - y_1^{[h]}$ denote the differences from the seed for data points in cluster h $(c_i = h)$
- Discarding the seed (marginalizing out to avoid order dependence), we define a partial likelihood contribution for cluster h: $\prod_{i=1}^{n_h} \prod_{j>i} g_h^{1/n_h}(\tilde{d}_{i,j}^{[h]})$
- $1/n_h$ is a calibration parameter & $g_h(.)$ is chosen to have expectation zero & symmetry with skewness zero
- In practice, we will often model pairwise distances $d_{i,j}^{[h]}$ instead of differences $\tilde{d}_{i,j}^{[h]}$ using a carefully chosen gamma kernel

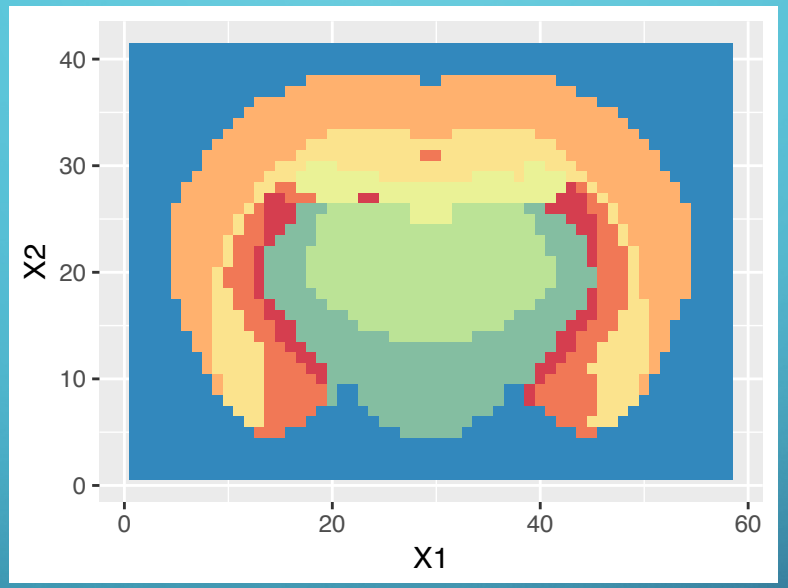
APPLICATION TO MOUSE BRAIN CLUSTERING

- Gene expression data from the Allen Mouse Brain Atlas (Lein et al 2007)
- For each voxel in a 41×58 region of a single mouse's brain, we get gene expression measurements for 3241 genes
- For illustration, we first extract the top 30 principal components
- We cluster these data and compare the clusters to known anatomical regions
- We applied both Bayesian distance clustering & a vanilla Gaussian mixture model (over-fitted vs of Rousseau & Mengersen 2011 $\rightarrow 7$ clusters)

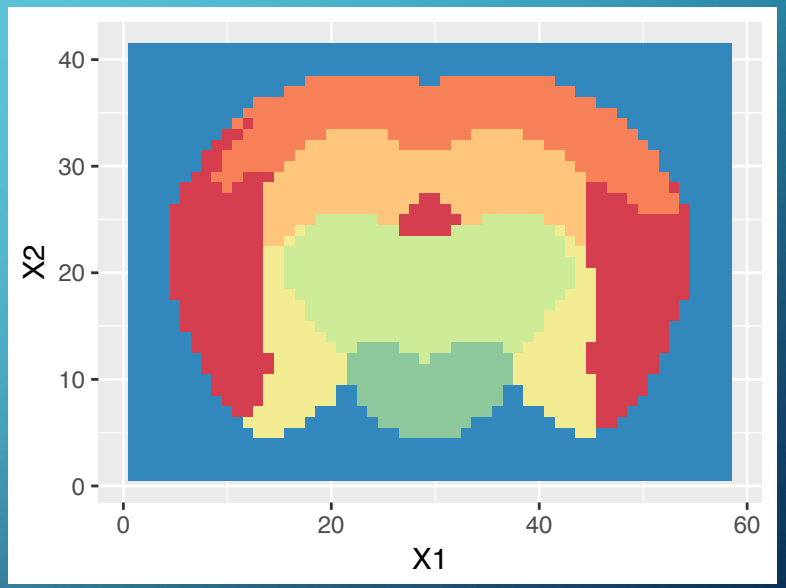
ANATOMICAL LABELS VS BDC & GMM



ANATOMICAL LABELS



Bayesian Distance Clustering



Clustering from Gaussian Mixture model

DISCUSSION – TOPIC 3

- Although clustering is a canonical topic in statistics & there has been a lot of work, a lot remains to be done in developing reliable & reproducible methods
- Bayesian nonparametric models tend to be heavily parametric in the clustering context & to inherit the disadvantages of parametric models
- Important to develop new inferential frameworks that seamlessly acknowledge M-open & model misspecification while remaining Bayesian
- Appealing to avoid ad hoc-ery in doing this & also not discard modeling entirely!

OVERALL DISCUSSION

- More broadly I think we need more original thinking in BNP work
- Modeling more & more complexity in data using increments on Dirichlet & Gaussian processes (etc) only gets you so far
- Real world data in many applications are more complex than our models can currently handle – practitioners have moved toward black box neural nets
- Can we embrace the complexity & challenges of real world applications & develop actually state-of-the-art + interpretable methods?
- If not we run the risk of becoming a narrow niche area