STATS305C: Applied Statistics III

Lecture 19: Wrapping Up

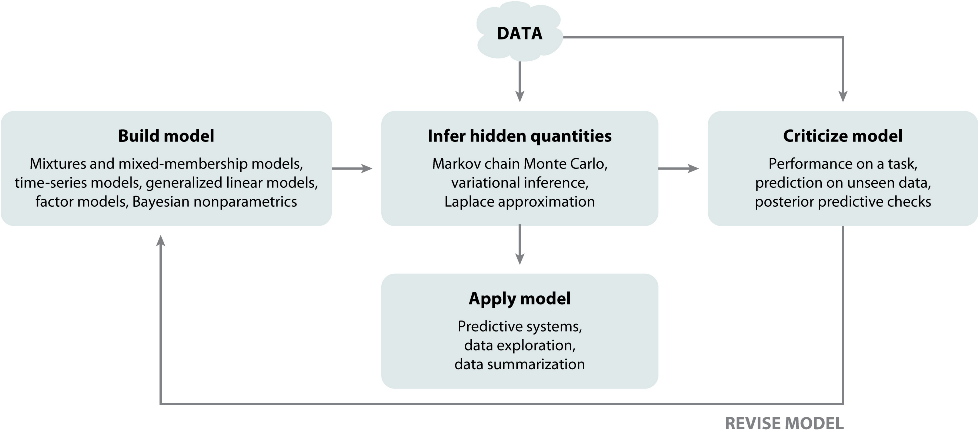
Scott Linderman

June 1, 2022

Recap

Model	Algorithm	Application
Multivariate Normal Models	Conjugate Inference	Bayesian Linear Regression
Hierarchical Models	MCMC (MH & Gibbs)	Modeling Polling Data
Probabilistic PCA & Factor Analysis	MCMC (HMC)	Images Reconstruction
Mixture Models	Expectation Maximization	Image Segmentation
Mixed Membership Models	Coordinate Ascent VI	Topic Modeling
Variational Autoencoders	Black Box, Amortized VI	Image Generation
State Space Models	Message Passing	Segmenting Video Data
Stochastic Processes	MCMC & Data Augmentation	Inhomog. Poisson Processes

Box's Loop



DATA

Build model

Mixtures and mixed-membership models, time-series models, generalized linear models, factor models, Bayesian nonparametrics

Infer hidden quantities

Markov chain Monte Carlo, variational inference, Laplace approximation

Criticize model

Performance on a task, prediction on unseen data, posterior predictive checks

Apply model

Predictive systems, data exploration, data summarization

REVISE MODEL

Blei DM. 2014.
Annu. Rev. Stat. Appl. 1:203–32

Blei [2014].

Outline

- **Bayesian model comparison**
- Posterior predictive checks

Marginal Likelihood

- The **marginal likelihood**, aka model evidence, is a useful measure of how well a model $\mathscr{M}_i$ fits the data.

- Specifically, it measures the *expected* probability assigned to the data under model $\mathscr{M}_i$, integrating over possible parameters under the prior,

$$
\begin{aligned}
p(\boldsymbol{x} \mid \mathscr{M}_i) &= \int p(\boldsymbol{\theta} \mid \mathscr{M}_i)\, p(\boldsymbol{x} \mid \boldsymbol{\theta}, \mathscr{M}_i)\, \mathrm{d}\boldsymbol{\theta} \qquad (1) \\
&= \mathbb{E}_{p(\boldsymbol{\theta} \mid \mathscr{M}_i)} \left[p(\boldsymbol{x} \mid \boldsymbol{\theta}, \mathscr{M}_i) \right] \qquad (2)
\end{aligned}
$$

- If a prior distribution puts high probability on parameters that then assign high conditional probability to the data, the marginal likelihood will be large.

- If the prior spreads its probability mass over a wide range of parameters, it may have a lower marginal likelihood than one that concentrates mass around the weights that achieve maximal likelihood.

Occam's Razor

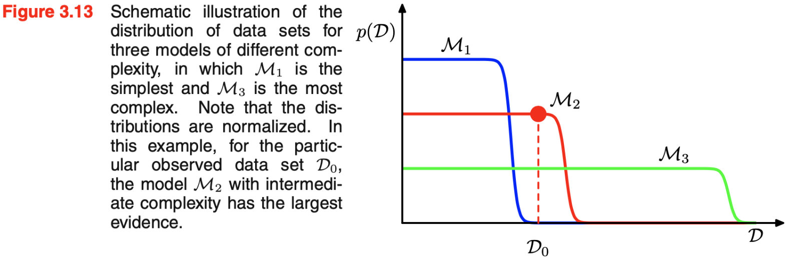


Figure 3.13 Schematic illustration of the distribution of data sets for three models of different complexity, in which $\mathcal{M}_1$ is the simplest and $\mathcal{M}_3$ is the most complex. Note that the distributions are normalized. In this example, for the particular observed data set $\mathcal{D}_0$, the model $\mathcal{M}_2$ with intermediate complexity has the largest evidence.

Figure: Bishop, pg 164

More flexible models can assign probability to many datasets, but since the distribution has to normalize, the probability of any given dataset is limited.

Thus, the marginal likelihood offers a form of **Occam's razor** for choosing models that are only as complex as is necessary.

Bayesian Model Averaging

Suppose you have a collection of models $\{\mathscr{M}_i\}_{i=1}^M$. How would a proper Bayesian leverage them to make predictions? *Put a prior on models and integrate over it!*

To make predictions, combine models according to their evidence,

$$p(x_{\text{new}} \mid \boldsymbol{x}) = \sum_{i=1}^{M} p(\mathscr{M}_i \mid \boldsymbol{x})\, p(x_{\text{new}} \mid \mathscr{M}_i) \tag{3}$$

$$\propto \sum_{i=1}^{M} p(\mathscr{M}_i)\, p(\boldsymbol{x} \mid \mathscr{M}_i)\, p(x_{\text{new}} \mid \mathscr{M}_i, \boldsymbol{x}) \tag{4}$$

A simple approximation is to make predictions using only the model with the highest evidence, $\mathscr{M}^\star$.

This is called **model selection**

Marginal likelihood in exponential family models

Recall that for exponential family distributions, the marginal likelihood is given by a ratio of normalizing constants,

$$p(\boldsymbol{x} \mid \mathscr{M}_i) = \left(\prod_{n=1}^{N} h(x_n) \right) \frac{Z(\boldsymbol{\phi}', \nu')}{Z(\boldsymbol{\phi}, \nu)} \tag{5}$$

where $\mathscr{M}_i$ is an exponential family model specified by prior hyperparameters $\boldsymbol{\phi}$ and ν.

The posterior parameters are,

$$\boldsymbol{\phi}' = \boldsymbol{\phi} + \sum_{n=1}^{N} t(x_n) \tag{6}$$

$$\nu' = \nu + N. \tag{7}$$

(We used these properties to derive collapsed Gibbs sampling algorithms last week.)

Example: Bayesian linear regression

In a Bayesian linear regression, the model is defined by the choice of features (basis functions) in the design matrix $\boldsymbol{X}$, as well as the prior hyperparameters $(\nu, \tau, \boldsymbol{\Lambda})$ of a normal-inverse-chi-squared prior,

$$
\begin{aligned}
p(\boldsymbol{y} \mid \boldsymbol{X}) = \int \frac{(2\pi)^{-\frac{N}{2}}}{Z(\nu, \tau^2, \boldsymbol{\Lambda})} (\sigma^2)^{-(1+\frac{\nu'}{2}+\frac{P}{2})} \\
\exp\Bigg\{ -\frac{1}{2}\Big\langle \nu' \tau'^2 + \boldsymbol{\mu}'^\top \boldsymbol{\Lambda}' \boldsymbol{\mu}', \frac{1}{\sigma^2} \Big\rangle \\
+ \Big\langle \boldsymbol{\Lambda}' \boldsymbol{\mu}', \frac{\boldsymbol{w}}{\sigma^2} \Big\rangle - \frac{1}{2}\Big\langle \boldsymbol{\Lambda}', \frac{\boldsymbol{w}\boldsymbol{w}^\top}{\sigma^2} \Big\rangle \Bigg\} \, \mathrm{d}\boldsymbol{w} \, \mathrm{d}\sigma^2 \qquad (8)
\end{aligned}
$$

$$
= (2\pi)^{-\frac{N}{2}} \frac{Z(\nu', \tau'^2, \boldsymbol{\Lambda}')}{Z(\nu, \tau^2, \boldsymbol{\Lambda})} \int \frac{1}{Z(\nu', \tau'^2, \boldsymbol{\Lambda}')} \text{``} \qquad \dots \qquad \text{''} \, \mathrm{d}\boldsymbol{w} \, \mathrm{d}\sigma^2 \qquad (9)
$$

$$
= (2\pi)^{-\frac{N}{2}} \frac{Z(\nu', \tau'^2, \boldsymbol{\Lambda}')}{Z(\nu, \tau, \boldsymbol{\Lambda})} \qquad (10)
$$

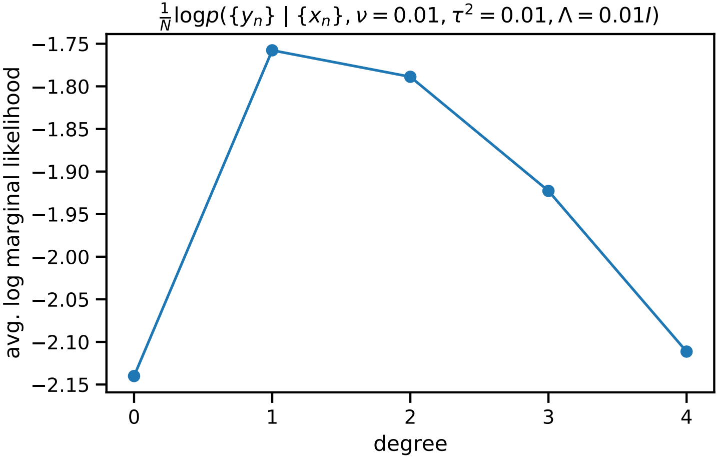
Example: Bayesian linear regression II

Under the conjugate prior, we can compute the marginal likelihood in closed form,

$$p(\boldsymbol{y} \mid \boldsymbol{X}) = (2\pi)^{-\frac{N}{2}} \frac{Z(\nu', \tau'^2, \boldsymbol{\Lambda}')}{Z(\nu, \tau, \boldsymbol{\Lambda})} \tag{11}$$

$$= (2\pi)^{-\frac{N}{2}} \frac{\Gamma(\frac{\nu'}{2})}{\Gamma(\frac{\nu}{2})} \frac{(\frac{\tau^2 \nu}{2})^{\frac{\nu}{2}}}{(\frac{\tau'^2 \nu'}{2})^{\frac{\nu'}{2}}} \frac{|\boldsymbol{\Lambda}|^{\frac{1}{2}}}{|\boldsymbol{\Lambda}'|^{\frac{1}{2}}} \tag{12}$$

Example: Bayesian linear regression III



The Evidence Approximation

- To get some insight into the model evidence, consider the case where $\theta \in \mathbb{R}$.
- Assume the posterior is peaked around its mode θ_{MAP} with width σ_{post}.
- Likewise, assume the prior is flat with width σ_{prior}.
- Then

$$p(\boldsymbol{x} \mid \mathscr{M}_i) = \int p(\boldsymbol{x} \mid \theta, \mathscr{M}_i)\, p(\theta \mid \mathscr{M}_i)\, \mathrm{d}\theta$$

$$\approx p(\boldsymbol{x} \mid \theta_{\text{MAP}} \frac{\sigma_{\text{post}}}{\sigma_{\text{prior}}}.$$

This is a simple *rectangular approximation.*

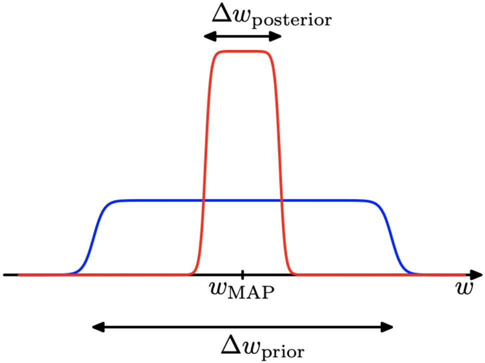


Figure: Bishop, pg. 163

Laplace Approximation

Idea: *approximate the posterior with a multivariate normal distribution centered on the mode.*

To motivate this, consider a second-order Taylor approximation to the log posterior,

$$\mathscr{L}(\boldsymbol{\theta}) \approx \mathscr{L}(\boldsymbol{\theta}_{\mathsf{MAP}}) + (\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathsf{MAP}})^\top \underbrace{\nabla_{\boldsymbol{\theta}} \mathscr{L}(\boldsymbol{\theta}_{\mathsf{MAP}})}_{\mathbf{0} \text{ at the mode}} + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathsf{MAP}})^\top \nabla^2_{\boldsymbol{\theta}} \mathscr{L}(\boldsymbol{\theta}_{\mathsf{MAP}})(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathsf{MAP}}) \tag{13}$$

$$= -\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathsf{MAP}})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta} - \boldsymbol{\theta}_{\mathsf{MAP}}) + c \tag{14}$$

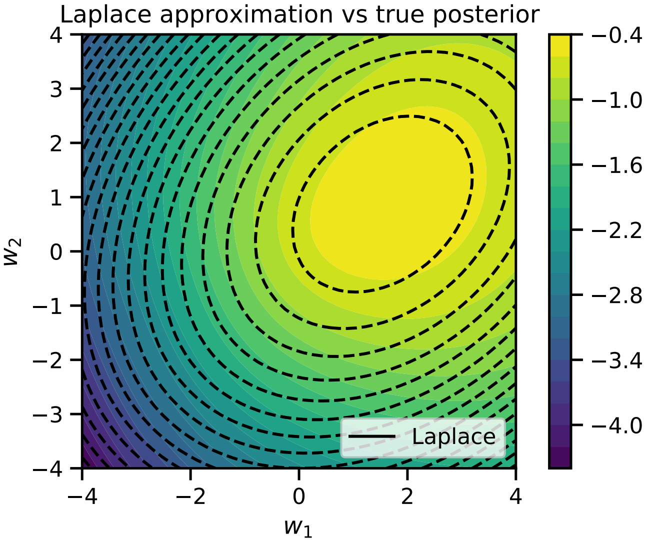
$$= \log \mathcal{N}(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\mathsf{MAP}}, \boldsymbol{\Sigma}) \tag{15}$$

where $\boldsymbol{\Sigma} = -[\nabla^2_{\boldsymbol{\theta}} \mathscr{L}(\boldsymbol{\theta}_{\mathsf{MAP}})]^{-1}$

In other words, the posterior is approximately Gaussian with covariance given by the (negative) inverse Hessian at the mode.

Since the Hessian is *negative* definite, the covariance is *positive* definite, as required.

Laplace Approximation II



Bernstein-von Mises Theorem

In the large data limit (as $N \to \infty$), the posterior is asymptotically normal, justifying the Laplace approximation in this regime.

Consider a simpler setting in which we have data $\{x_n\}_{n=1}^N \overset{\text{iid}}{\sim} p(x \mid \theta_{\text{true}})$.

Under some conditions (e.g. $\boldsymbol{\theta}_{\text{true}}$ not on the boundary of $\boldsymbol{\Theta}$ and $\boldsymbol{\theta}_{\text{true}}$ has nonzero prior probability), then the MAP estimate is consistent. As $N \to \infty$, $\boldsymbol{\theta}_{\text{MAP}} \to \boldsymbol{\theta}_{\text{true}}$.

Likewise,

$$p(\boldsymbol{\theta} \mid \{x_n\}_{n=1}^N) \to \mathcal{N}\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\text{true}}, \tfrac{1}{N}[J(\boldsymbol{\theta}_{\text{true}})]^{-1}\right) \tag{16}$$

where

$$J(\boldsymbol{\theta}) = -\mathbb{E}_{p(x \mid \boldsymbol{\theta})}\left[\frac{\mathrm{d}^2}{\mathrm{d}\boldsymbol{\theta}^2} \log p(x \mid \boldsymbol{\theta})\right] \tag{17}$$

is the *Fisher information* of parameter $\boldsymbol{\theta}$.

Approximating the marginal likelihood

The Laplace approximation also offers an approximation of the intractable marginal likelihood,

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\theta}_{\mathsf{MAP}}) &= \log p(\boldsymbol{\theta}_{\mathsf{MAP}} \mid \mathcal{M}_i) + \log p(\boldsymbol{x} \mid \boldsymbol{\theta}_{\mathsf{MAP}}, \mathcal{M}_i) - \log p(\boldsymbol{x} \mid \mathcal{M}_i) \qquad (18)\\
&\approx \log \mathcal{N}(\boldsymbol{\theta}_{\mathsf{MAP}} \mid \boldsymbol{\theta}_{\mathsf{MAP}}, \boldsymbol{\Sigma}) \qquad (19)\\
&= -\frac{P}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}| \qquad (20)
\end{aligned}$$

where again, $\boldsymbol{\Sigma} = -[\nabla^2_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}_{\mathsf{MAP}})]^{-1}$. Rearranging terms,

$$\log p(\boldsymbol{x} \mid \mathcal{M}_i) \approx \log p(\boldsymbol{\theta}_{\mathsf{MAP}}) + \log p(\boldsymbol{x} \mid \boldsymbol{\theta}_{\mathsf{MAP}}, \mathcal{M}_i) + \frac{D}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Sigma}|$$

Combine this with $\boldsymbol{\Sigma} \approx \frac{1}{N}[J(\boldsymbol{\theta}_{MAP})]^{-1}$ and $\frac{1}{2}\log|\boldsymbol{\Sigma}| \approx \frac{D}{2}\log N + O(1)$ to derive the **Bayesian information criterion (BIC)**, a technique for penalized maximum likelihood estimation.

Approximating the marginal likelihood with importance sampling

- We can obtain an unbiased estimate of the marginal likelihood with ordinary Monte Carlo,

$$p(\boldsymbol{x} \mid \mathscr{M}_i) = \int p(\boldsymbol{x} \mid \boldsymbol{\theta}, \mathscr{M}_i)\, p(\boldsymbol{\theta} \mid \mathscr{M}_i)\, \mathrm{d}\boldsymbol{\theta} \tag{21}$$

$$\approx \frac{1}{S} \sum_{s=1}^{S} p(\boldsymbol{x} \mid \boldsymbol{\theta}^{(s)}, \mathscr{M}_i) \qquad \boldsymbol{\theta}^{(s)} \overset{\text{iid}}{\sim} p(\boldsymbol{\theta} \mid \mathscr{M}_i) \tag{22}$$

  but these estimates are often **exceedingly high variance**.

- It would be better if we could target our samples toward regions that have high likelihood. **Importance sampling** aims to do that via a **proposal distribution** $r(\boldsymbol{\theta})$,

$$p(\boldsymbol{x} \mid \mathscr{M}_i) \approx \frac{1}{S} \sum_{s=1}^{S} w^{(s)}\, p(\boldsymbol{x} \mid \boldsymbol{\theta}^{(s)}, \mathscr{M}_i) \qquad \boldsymbol{\theta}^{(s)} \overset{\text{iid}}{\sim} r(\boldsymbol{\theta}) \tag{23}$$

  where $w^{(s)} \triangleq \frac{p(\boldsymbol{\theta}^{(s)} \mid \mathscr{M}_i)}{r(\boldsymbol{\theta}^{(s)})}$ is the **importance weight**.

Importance sampling II

- Ideally, we would propose from the posterior distribution $r(\boldsymbol{\theta}) = p(\boldsymbol{\theta} \mid \boldsymbol{x}, \mathscr{M}_i)$.
- Then the estimator would have zero variance since,

$$p(\boldsymbol{x} \mid \mathscr{M}_i) \approx \frac{1}{S} \sum_{s=1}^{S} \frac{p(\boldsymbol{\theta}^{(s)} \mid \mathscr{M}_i)\, p(\boldsymbol{x} \mid \boldsymbol{\theta}^{(s)}, \mathscr{M}_i)}{p(\boldsymbol{\theta}^{(s)} \mid \boldsymbol{x}, \mathscr{M}_i)}) \tag{24}$$

$$= \frac{1}{S} \sum_{s=1}^{S} p(\boldsymbol{x} \mid \mathscr{M}_i) \tag{25}$$

$$= p(\boldsymbol{x} \mid \mathscr{M}_i) \tag{26}$$

- Of course, we can't sample the posterior exactly for the model's we're interested in here!

Annealed Importance Sampling

- Annealed importance sampling [Neal, 2001] is a way of constructing a proposal distribution by sampling a sequence of parameter values $\boldsymbol{\theta}_T, \ldots, \boldsymbol{\theta}_0$, where $\boldsymbol{\theta}_0 \equiv \boldsymbol{\theta}$ is our final proposal.

- The idea is to set,

$$r(\boldsymbol{\theta}_0) = \int r(\boldsymbol{\theta}_T, \ldots, \boldsymbol{\theta}_0) \, \mathrm{d}\boldsymbol{\theta}_{T:1} \tag{27}$$

  where

$$r(\boldsymbol{\theta}_T, \ldots, \boldsymbol{\theta}_0) = r_T(\boldsymbol{\theta}_T) \, r_{T-1}(\boldsymbol{\theta}_{T-1} \mid \boldsymbol{\theta}_T) \cdots r_0(\boldsymbol{\theta}_0 \mid \boldsymbol{\theta}_1). \tag{28}$$

- We choose the sequence of conditional distributions $r_t(\boldsymbol{\theta}_t \mid \boldsymbol{\theta}_{t+1})$ to be **Markov transition operators** with stationary distributions $f_t(\boldsymbol{\theta}_t)$ that **anneal** from the prior $f_T(\boldsymbol{\theta}_T) = p(\boldsymbol{\theta}_T \mid \mathscr{M}_i)$ to the posterior $f_0(\boldsymbol{\theta}_0) = p(\boldsymbol{\theta}_0 \mid \boldsymbol{x}, \mathscr{M}_i)$.

- For example, $f_t(\boldsymbol{\theta}_t) \propto p(\boldsymbol{\theta}_t \mid \mathscr{M}_i) \, p(\boldsymbol{x} \mid \boldsymbol{\theta}_t, \mathscr{M}_i)^{\beta_t}$ for $\beta_T = 0 < \beta_{T-1} \cdots < \beta_0 = 1$.

Empirical Bayes and Type-II Maximum Likelihood

- What about the hyperparameters $\boldsymbol{\phi}$ and ν that define model $\mathscr{M}_i$?

- If we were super-duper Bayesian, we would put a prior on our prior hyperparameters and integrate over them, but that just kicks the can down the road. At some point we need to commit...

- **Empirical Bayes**, a.k.a. **type-II maximum likelihood estimation**, use point estimates of the hyperparameters chosen in a data-dependent manner,

$$\begin{aligned} \boldsymbol{\phi}^*, \nu^* &= \arg\max p(\boldsymbol{x} \mid \boldsymbol{\phi}, \nu) && (29) \\ &= \arg\max \int p(\boldsymbol{x} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta} \mid \boldsymbol{\phi}, \nu)\, \mathrm{d}\boldsymbol{\theta}. && (30) \end{aligned}$$

- For exponential families, the objective can be computed in closed form; for more complex models, approximations like the Laplace approximation can be used instead.

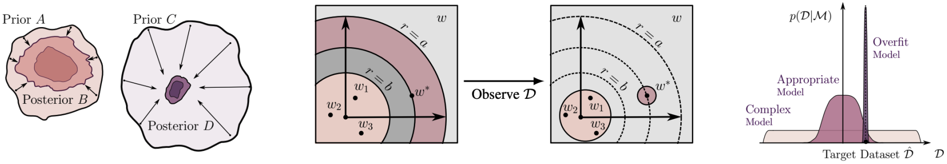
- In either case, the optimal hyperparameters typically need to be found via generic optimization algorithms like gradient descent.

Caveats...

- Note that in order for the marginal likelihood to be meaningful, we need to have a **proper prior** distribution. In the uninformative/improper limit, the marginal likelihood goes to zero.
- Bayesian model selection based on the marginal likelihood only really makes sense when we have a **finite set of models** $\{\mathscr{M}_i\}$**.**
- The marginal likelihood **does not measure generalization**. It measures the expected probability of the *observed data under the prior*, not the expected probability of *new data under the posterior*.

Current research

Bayesian model comparison, marginal likelihood estimation, and generalization are still topics of research, especially as [Lotfi et al., 2022].



(a) Posterior contraction (b) Marginal likelihood underfitting (c) Marginal likelihood overfitting

Figure 1. **Pitfalls of marginal likelihood.** **(a)**: Prior B is vague, but contains easily identifiable solutions and quickly collapses to posterior D after observing a small number of datapoints. Prior A describes the data better than prior B, but posterior D describes the data better than posterior B. The marginal likelihood will prefer model A, but model C generalizes better. **(b)**: Example of misalignment between marginal likelihood and generalization. The marginal likelihood will pick prior scale b, and not include the best solution w^*, leading to suboptimal generalization performance. **(c)**: The complex model spreads its mass thinly on a broad support, while the appropriate model concentrates its mass on a particular class of problems. The overfit model is a δ-distribution on the target dataset $\hat{\mathcal{D}}$.

Figure: From Lotfi et al. [2022]

Outline

- Bayesian model comparison
- **Posterior predictive checks**

Posterior Predictive Distribution

- One of the main uses of regression models is to make predictions, e.g. of y_{N+1} at $\boldsymbol{x}_{N+1}$.
- In Bayesian data analysis, this is given by the *posterior predictive distribution*,

$$p(y_{N+1} \mid \boldsymbol{x}_{N+1}, \{y_n, \boldsymbol{x}_n\})_{n=1}^{N}) = \int p(y_{N+1} \mid \boldsymbol{x}_{N+1}, \boldsymbol{w}, \sigma^2)\, p(\boldsymbol{w}, \sigma^2 \mid \{y_n, \boldsymbol{x}_n\}_{n=1}^{N}\, \mathrm{d}\boldsymbol{w}\, \mathrm{d}\sigma^2 \qquad (31)$$

- Generally, we can approximate the posterior predictive distribution with Monte Carlo.
- For Bayesian linear regression with a conjugate prior, we can compute it in closed form.

Model checking

The following slides are adapted from Aki Vehtari's lecture notes.

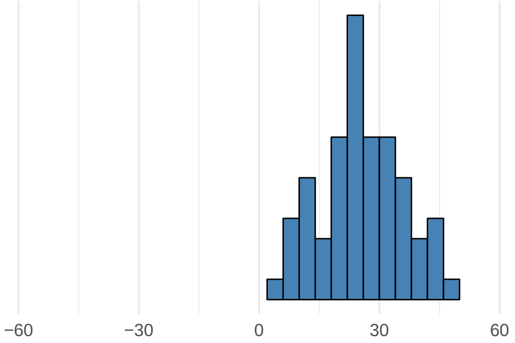
https://github.com/avehtari/BDA_course_Aalto/blob/master/slides/

Posterior predictive checks (PPCs)

- Newcomb's speed of light measurements
- Model:

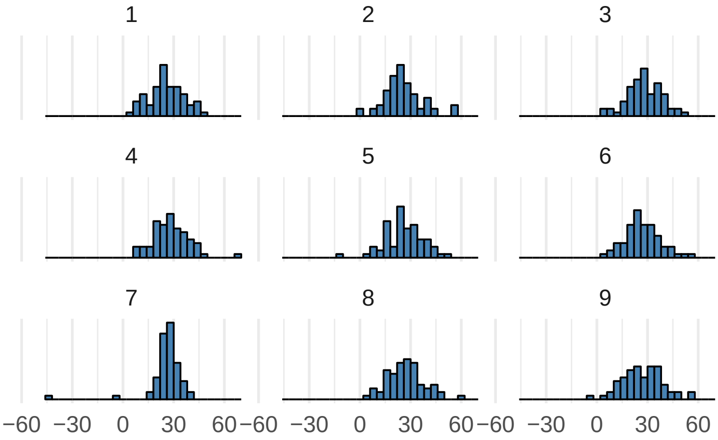
$$y \sim \mathcal{N}(\mu, \sigma)$$
$$p(\mu, \log \sigma) \propto 1$$

- Posterior predictive replicate y^{rep}
  - draw $\mu^{(s)}, \sigma^{(s)}$ from the posterior $p(\mu, \sigma \mid y)$
  - draw $y^{\mathrm{rep}\,(s)}$ from $\mathcal{N}(\mu^{(s)}, \sigma^{(s)})$
  - repeat n times to get y^{rep} with n replicates
- y^{rep} refers to replicating the whole experiment (potentially with same values of x) and obtaining as many replicated observations as in the original data.



Posterior predictive checks (PPCs) II

- Generate several replicated datasets y^{rep}
- Compare to the original dataset

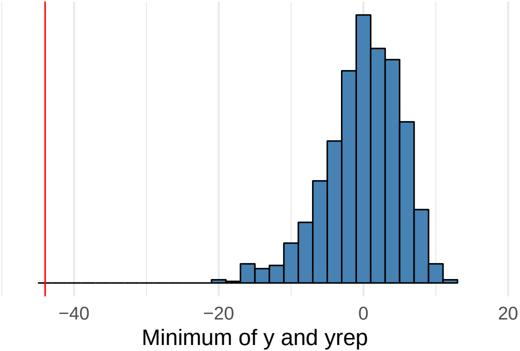


Posterior predictive checking with test statistic

- Replicated data sets y^{rep}
- Test quantity (or discrepancy measure) $T(y, \theta)$
    - summary quantity for the observed data $T(y, \theta)$
    - summary quantity for a replicated data $T(y^{\mathrm{rep}}, \theta)$
    - can be easier to compare summary quantities than data sets

Example: Posterior predictive checking with the min

- Compute test statistic for data $T(y, \theta) = \min(y)$
- Compute test statistic $\min(y^{\text{rep}})$ for many replicated datasets



Posterior predictive checking

- *Posterior predictive p-value*

  $$\begin{aligned} p &= \Pr(T(y^{\mathrm{rep}}, \theta) \geq T(y, \theta) \mid y) \\ &= \int\int \mathbb{I}[T(y^{\mathrm{rep}}, \theta) \geq T(y, \theta)]\, p(y^{\mathrm{rep}} \mid \theta) p(\theta \mid y)\, \mathrm{d}y^{\mathrm{rep}}\, \mathrm{d}\theta \end{aligned}$$

  where I is an indicator function

  - having $(y^{\mathrm{rep}\,(s)}, \theta^{(s)})$ from the posterior predictive distribution, easy to compute

    $$T(y^{\mathrm{rep}(s)}, \theta^{(s)}) \geq T(y, \theta^{(s)}), \quad s = 1, \ldots, S$$

- Posterior predictive p-value (ppp-value) estimated whether difference between the model and data could arise by chance
- Not commonly used, since the distribution of test statistic has more information

Sensitivity analysis

- How much different choices in model structure and priors affect the results
  - test different models and priors
  - alternatively combine different models to one model
    - e.g. hierarchical model instead of separate and pooled
    - e.g. *t* distribution contains Gaussian as a special case
  - robust models are good for testing sensitivity to "outliers"
    - e.g. *t* instead of Gaussian
- Compare sensitivity of essential inference quantities
  - extreme quantiles are more sensitive than means and medians
  - extrapolation is more sensitive than interpolation

What would I cover if I had 10 more weeks?

- More state space models!
  - Switching linear dynamical systems (SLDS) and "Recurrent" SLDS
  - Sequential VAEs, structured VAEs, and deep state space models
- Sequential Monte Carlo methods
- Disentangling and identifiability in nonlinear latent variable models
- Bayesian deep learning
- More Monte Carlo methods: slice sampling, NUTS, quasi-Monte Carlo, ...
- Undirected graphical models, energy based models, contrastive divergence, score matching...
- Density ratio estimation
- Suggestions?

Recap

Model	Algorithm	Application
Multivariate Normal Models	Conjugate Inference	Bayesian Linear Regression
Hierarchical Models	MCMC (MH & Gibbs)	Modeling Polling Data
Probabilistic PCA & Factor Analysis	MCMC (HMC)	Images Reconstruction
Mixture Models	Expectation Maximization	Image Segmentation
Mixed Membership Models	Coordinate Ascent VI	Topic Modeling
Variational Autoencoders	Black Box, Amortized VI	Image Generation
State Space Models	Message Passing	Segmenting Video Data
Stochastic Processes	MCMC & Data Augmentation	Inhomog. Poisson Processes

The End

Thank you all for a wonderful quarter, and have a great summer!

Please take time to fill out the course evaluation so I can improve for next year.

References I

David M Blei. Build, compute, critique, repeat: Data analysis with latent variable models. *Annu. Rev. Stat. Appl.*, 1(1):203–232, January 2014.

Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11(2):125–139, 2001.

Sanae Lotfi, Pavel Izmailov, Gregory Benton, Micah Goldblum, and Andrew Gordon Wilson. Bayesian model selection, the marginal likelihood, and generalization. *arXiv preprint arXiv:2202.11678*, 2022.