# What to know for midterm #1

## 1 BDA 4: Asymptotics and connections to non-Bayesian approaches

### 1.1 Consistency

- Know the definition of what it means for an estimator to be consistent.
- Know the definition of what it means for a posterior distribution to be consistent.
- Understand the concept of frequentist analysis of Bayesian methods, in which we ask what properties a given Bayesian procedure would have if the data were generated from some true distribution $P_0$ (which may or may not be a member of the assumed model class).
- Know that Doob's theorem provides very general conditions under which posterior consistency holds, if the model is correct specified and identifiable. (You do not need to know the details.)
- Know that when the model is misspecified, the posterior will typically concentrate at the point $\theta^*$ minimizing the KL divergence.

### 1.2 Asymptotic normality

- Know what it means that the posterior is asymptotically normal, and be able to write down the formula expressing this. In particular, know the mean and covariance matrix of the normal approximation.
- Be able to derive the formula for asymptotic normality of the posterior, from the Taylor approximation (without rigorous details).

- In simple cases with univariate $\theta$, be able to analytically compute the mean and variance of the asymptotic normal approximation, for a given likelihood.
- (Exercise 4 from homework 1) Understand why the posterior on $\phi$ is (typically) asymptotically normal when $\phi = f(\theta)$, and know how the mean and variance of the asymptotic normal distribution change under this transformation.
- Be able to give an argument for why asymptotic normality of the posterior, plus consistency of the MLE, typically will imply posterior consistency.
- Understand some of the ways in which posterior consistency and asymptotic normality can fail, and be able to give examples.

## 1.3 Frequentist coverage

- Know the definition of the coverage probability of a confidence region, and have a good intuitive understanding of what it means.
- Understand why having good frequentist coverage is a desirable property.
- In simple cases, be able to analytically compute coverage probability.
- Understand the definition of a posterior credible region (e.g., a 90% or 95% credible region).
- Know the definition of an equal-tailed posterior credible interval.
- In simple cases, be able to analytically compute an equal-tailed posterior credible interval.
- Know that posterior credible regions often (but not always) have good frequentist coverage properties.
- Understand why, if the prior and likelihood are exactly correct, posterior credible regions have frequentist coverage equal to their posterior probability (exercise 15a from homework 1).

# 2 BDA 6 & 7: Model checking and cross-validation

## 2.1 Posterior predictive checking

- Understand the idea behind posterior predictive checks.
- Know how to perform a posterior predictive check and compute a posterior predictive p-value.
- In simple cases, be able to analytically compute a posterior predictive p-value.
- Know the definition of the posterior predictive distribution for replicate data sets, and know how to sample from it based on posterior samples.
- Know how to interpret the results of a posterior predictive check.
- Understand that, ideally, p-values are uniformly distributed, so sometimes we will see p-values close to zero or one simply by chance. If we were to compute a large number of p-values, know how many we would expect to see outside a given range.
- Know that (unfortunately) posterior predictive p-values are not "true" p-values in the sense that they are not uniformly distributed, even if the model is correct.
- Understand why some test statistics/quantities will always be well captured by a given model (and thus are not very informative about model fit), especially in the case of exponential families.
- Realize that one needs to be careful when modifying the model based on the results of posterior predictive checks, since this can lead to overfitting.
- Realize that posterior predictive checks represent a sort of internal consistency check, but that they are not an ideal way of evaluating model fit, because they are "using the data twice".

## 2.2 Cross-validation

- Understand the idea behind cross-validation—why does it make sense?
- Know the definition of leave-one-out cross-validation.
- Know the definition of $k$-fold cross-validation.
- Know these common choices of loss function for cross-validation: log posterior predictive, 0-1 loss, square loss.
- Be able to derive the expected loss for a given loss function (taking care to compute it with respect to the true distribution!)
- Know how to compute a cross-validation estimate of generalization performance.
- Understand why cross-validation will typically provide a better assessment of performance compared to posterior predictive checks, since CV is not evaluating the model on the same data that was used to fit the model.
- Understand that if cross-validation is used to choose among multiple models, then in order to assess the performance of the chosen model, it needs to be evaluated on a further held-out set (disjoint from the set of data used for cross-validation).

# 3 BDA 8: Modeling accounting for data collection

- Be able to recognize situations in which the data collection process is biased in a way that will affect your inferences.
- Be able to give specific examples of situations in which it is important to model the data collection process.
- Know the definition of ignorability, as well as the intuitive interpretation of it.
- Be able to explain the interpretation of the "potential outcomes" $y$ and the observation indicators $I$.

- Know the definition of the “complete-data likelihood” in the general setup we considered.
- Know what distribution to use for posterior inferences about $\theta$ when ignorability does not hold.
- Be able to derive a formula for this posterior in simple cases.
- Given a verbal description of a distribution on potential outcomes and a data collection process, be able to write down a reasonably appropriate probabilistic model for it.
- Know the definitions of the following conditions: missing at random (MAR), missing completely at random (MCAR), strong ignorability, and distinct parameters.
- Be able to give examples in which these conditions hold or do not hold.
- Know what implications hold between these different conditions.
- Be able to prove that strong ignorability implies ignorability.
- Be able to prove that MAR + distinct parameters implies ignorability.