Posterior Predictive Checks

David M. Blei
Princeton University

December 16, 2011

1 Motivation

- No model is correct—all models are an approximation (Box).

- Some models aren't even that. Many models are convenient formalisms for specifying hidden structure that we want to uncover to form hypotheses or make predictions.

- Model building involves three components: estimation, criticism, revision.
  - Estimation: Estimate (or infer) parameters conditioned on the truth of the model.
  - Criticism "involves a confrontation of [the model] with available data (old as well as newly acquired) and asks whther [the model] is consonant with [it]" (Box, 1980).
  - Revision decides how to change the model based on the criticisms (and available time/computing/knowledge resources).

  We know how to do estimation. We need guidance about criticism and revision.

- Today we discuss *criticism*.
  - Is my model good enough?
  - And in what ways isn't it?

- This is the art (or mud) of statistics. These questions are vague.

- Note that this is different from *model comparison*. We are not choosing among several models, but deciding whether and in what way to change a model.

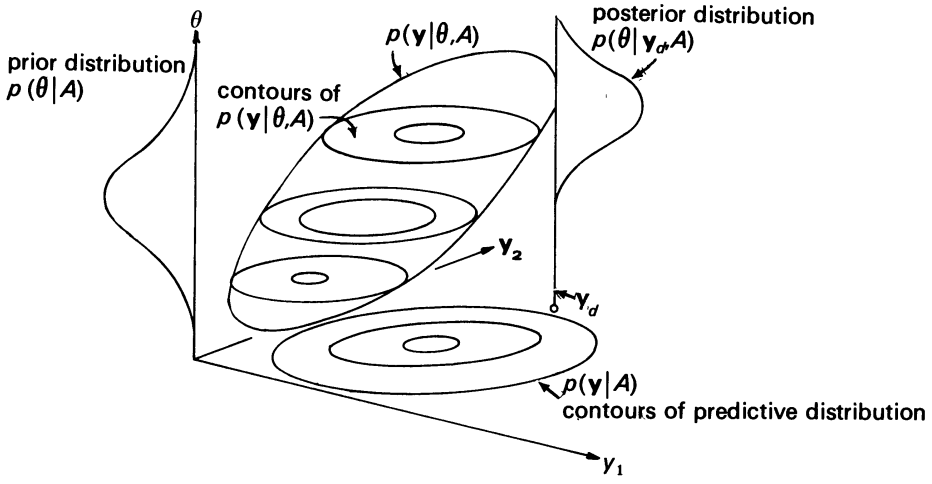
- This feels even more relevant today. I think of modeling as piecing together various modules, rather than choosing among a population of models.
- Machine learning has given us many new building blocks, but has little to say about how to diagnose models.
- This is especially important in exploratory analysis, e.g., to form hypotheses or organize data. Many exploratory tasks do not have clear measures of quality.

- Automating model building is a tall order. Even BNP methods do not automate it.
  - They help define flexible models, but it is up to the modeler to define likelihood functions, dependencies between the observed data and latent variables, etc.

2 The predictive check

- Box (1980) describes a *predictive check*, which tells the story. (Though this story will be refined in a posterior predictive check.)

- All the intuitions about how to assess a model are in this picture:



- The set up from Box (1980) is the following.
  - The data are $\mathbf{y}$; the hidden variables are θ; the model is M.
  - Each point of the hidden variable θ yields a distribution of data.
  - The joint distribution combines the prior and the likelihood

$$p(y,\theta\,|\,M) = p(\mathbf{y}\,|\,\theta)p(\theta\,|\,M) \tag{1}$$

- But it can also be factored as

$$p(\mathbf{y}, \theta \mid M) = p(\theta \mid \mathbf{y}, M) p(\mathbf{y} \mid M). \tag{2}$$

- "Prior" predictive checking:
  - Suppose we observe a data set $\mathbf{y}_d$.
  - The first factor is the posterior of the hidden variables. If the model is believed then it allows "all relevant estimation inferences to be made about θ."
  - But suppose the data is unlikely to have been generated by this model, that is, suppose the model is suspect. This cannot be detected from the posterior.
  - The second factor is the *predictive distribution*,

$$p(y \mid M) = \int_\theta p(y \mid \theta) p(\theta \mid M) \tag{3}$$

  - This is the *distribution of data sets if the model is true*.
  - If the model is suspect, then locating $\mathbf{y}_d$ in $p(\mathbf{y} \mid M)$ will reveal that the observed data has low probability under the assumed model.
  - Or reference a checking function $g(\mathbf{y}_d)$ in its probability under the marginal.

- We are going to change this a bit, but the main ideas are here.
  - Gelman: "If an inadequate model is fit, we can get precise, but wrong, inferences."
  - We can design our model to be able to compute the posterior; we can design our function $g(\mathbf{y})$ to reflect what we care about the model being right about.
  - We compare observations against a reference distribution. (It's frequentist!)

- Box (1980) sees this as blending Bayesian and Frequentist ideas:

> In [model building] many subjective choices are made, concious or unconcious, good or bad...The wisdom of these choices over successive stages of development is the major determinant of how fast the iteration will converge or of whether it converges at all, and distinguishes good scientists and statisticians from bad. It is in this context that theories of inference need to be considered.

- In particular,

> [Frequentist] sampling theory is needed for exploration and ultimate *criticism* of an entertained model in light of current data, while Bayes' theory is needed for *estimation* of parameters conditional on the adequacy of the entertained model.

3 The posterior predictive check

- Devised in Rubin (1984). Expanded in Gelman et al. (1996).

- Compares observed data to a reference distribution based on the posterior (conditioned on the observed data).

- Rubin summarizes—

  > Given observed data, X_{obs}, what would we expect to see in hypothetical replications of the study that generated X_{obs}? Intuitively, if the model specifications are appropriate, we would expect to see something similar to what we saw this time, at least similar in "relevant ways." This statement, which is essentially a fundamental premise of frequency inference seems to me so basic that it needs no defense. Unlike the frequentist, the Bayesian, though, will condition on all observed values.

  Notes:

  - In frequentist estimation, such checks are often done with the MLE. This too "conditions" on the observed values.
  - We can also play the same game with future data. How well to replicated data match real future data? (We'll discuss this too.)

- Continuing to quote Rubin. (Bullets and emphasis are mine.)

  - In order to apply the idea, we first need to define a statistic T(X) that formalizes the notation of "relevant ways."
  - We need to define precisely what we mean by a *replication of the current study*.
  - Having defined $T(X)$ and the replications, we then calculate the frequency distribution of $T(X)$ in the hypothetical future replications, where this distribution is conditional on both (a) the observed data X_{obs} and (b) the current model specification $f(X|\theta)p(\theta)$. [DB: Box (1980) only conditions on the model.]
  - This distribution—the model monitoring distribution or posterior predictive check distribution—is the posterior predictive distribution of $T(X)$, "posterior" meaning conditional on observed values and "predictive" meaning the distribution of a future observable quantity.
  - If the frequency distribution of $T(X)$ does not make the observed value of $T(X)$, $T(X_{\text{obs}})$ appear typical, where typical is usually defined by tail areas of the distribution of $T(X)$ beyond $T(X_{\text{obs}})$, then we may want to revise the model $f(X|\theta)p(\theta)$. The reason is that the model, in replications of the current study, does not generate data that are similar to the observed data, where similar is judged by comparing $T(X_{\text{obs}})$ to the distribution of $T(X)$.

- Now we've added two more arts to the art of model building—designing a discrepancy and defining replications.

- Paraphrasing Rubin: Designed to reveal a lack of fit, rather than being sufficient in an expanded model. He makes the further point that building increasingly complex models and then stopping one step beyond when you would stop is too much work.

4 Details (from GMS)

- Let $\mathbf{y}$ be the observed data; H be the assumed model; θ be an unknown variable/parameters.

- Define $\mathbf{y}^{\text{rep}}$ to be the replicated data—data that could have been observed if the same experiment was repeated with the same model and the same (unknown) value of θ that produced $\mathbf{y}$.

- This replication has distribution $P(\mathbf{y}^{\text{rep}} \mid H, \theta)$.

- A classical p-value based on a discrepancy T is

$$p_c(\mathbf{y}, \theta) = P(T(\mathbf{y}^{\text{rep}}) \geq T(\mathbf{y}) \mid H, \theta). \tag{4}$$

  - The only thing random here is $\mathbf{y}^{\text{rep}}$.
  - This value is obtainable when the distribution of p_c does not depend on θ.
  - E.g., in linear regression, we can define T to be a standardized residiual (or something like that) and know the distribution of $T(\mathbf{y}^{\text{rep}})$ and p_c.
  - What is important is that this let's us locate $\mathbf{y}$ in the distribution of $\mathbf{y}^{\text{rep}}$.

- Consider a Bayesian set-up, where there is a distribution of θ.

  - The reference distribution of future observations is

$$P(\mathbf{y}^{\text{rep}} \mid H, y) = \int P(\mathbf{y}^{\text{rep}} \mid H, \theta) p(\theta \mid H, \mathbf{y}) d\theta \tag{5}$$

  - We then can plot the observed value of $T(\mathbf{y})$ against the distribution of $T(\mathbf{y}^{\text{rep}})$. The tail probability is

$$\begin{aligned} p_b(y) &= P(T(\mathbf{y}^{\text{rep}}) \geq T(\mathbf{y}) \mid H, y) = \int p_c(y, \theta) P(\theta \mid H, \mathbf{y}) d\theta & (6) \\ &= \int P(T(\mathbf{y}^{\text{rep}}) \geq T(\mathbf{y}) \mid H, \theta) P(\theta \mid H, \mathbf{y}) d\theta. & (7) \end{aligned}$$

- This is the classical p-value averaged over the posterior. This is the *posterior predictive p-value*. Compare to Box—he averages over the prior.
- Look at where $\mathbf{y}^{\text{rep}}$ comes from. It is drawn conditioned on θ; θ is drawn conditioned on observations $\mathbf{y}$. Thus, $\mathbf{y}^{\text{rep}}$ comes from the posterior predictive distribution—it is from the distribution of new data.

- These schematics illustrate the difference.

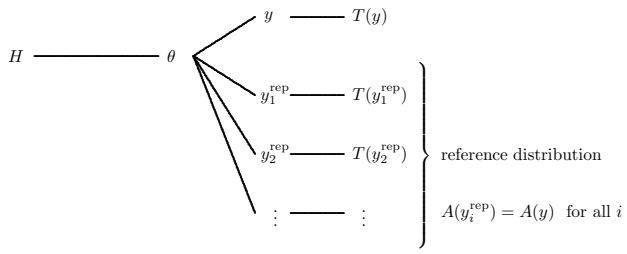


Figure 1a. The posterior predictive distribution

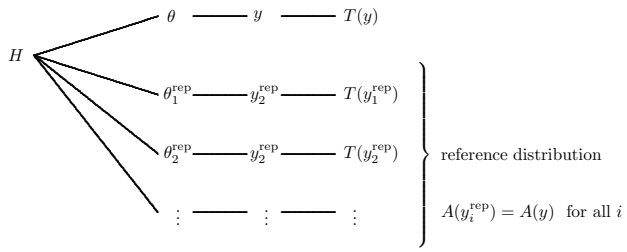


Figure 1b. The prior predictive distribution

- The top plot is the classical goodness-of-fit test. The variable θ is the MLE and we check the observations against the reference distribution given by θ.
- Note how GOF tests work. We fit the model to data and compute the standardized residuals. Its distribution does not depend on θ—that is the design principle.
- In the PPC, we use the top plot again, but marginalize out over the prior of θ. *This let's us consider any discrepancy as long as we can simulate from the posterior.* We are no longer limited to discrepancies for which we know the distribution.
- The bottom plot is Box's set-up. It is different from a GOF test. It is less intuitive because it checks the prior (i.e., the model parameters).
- Gelman et al. says: PPCs treat the prior as an outmoded first guess; Box treats it as the true population distribution.
- Rubin weighs in: "For example, in a study of D drugs, interest may focus on the fit of the model for these drugs rather than for a sample of D new drugs rawn from the same population of drugs."
- (But there are criticisms of this too. I think there is middle ground.)

- GMS's innovation lets the discrepancy function depend on the hidden variables, $D(\mathbf{y};\theta)$.
  - It then has a reference distribution

$$P(\mathbf{y}^{\text{rep}},\theta\,|\,H,\mathbf{y}) = p(\mathbf{y}^{\text{rep}}\,|\,H,\theta)p(\theta\,|\,\mathbf{y}) \tag{8}$$

    Note the earlier reference distribution is the marginal of this.
  - The tail probability is defined in the same way

$$p_b(\mathbf{y}) = P(D(\mathbf{y}^{\text{rep}};\theta) \geq D(\mathbf{y};\theta)\,|\,H,\mathbf{y}) \tag{9}$$

  - (This includes the old p_b as a special case.)

- This discrepancy is naturally handled via simulation
  - Simulate θ^b from the posterior (e.g., with MCMC)
  - Simulate a replicated data set $\mathbf{y}^{\text{rep}}$ from θ^b.
    - Footnote: Rubin indicates that the replication can depend on θ.
  - Compute $D(\mathbf{y}^{\text{rep}};\theta^b)$ and $D(\mathbf{y};\theta^b)$

- Make the scatter plot; the proportion about the 45 degree line is the p-value.
  - Plots are the most useful diagnostic tool. In later papers, Gelman notes that classical diagnostic plots are also about comparing observations to a reference.

5 Wrap-up

- Appealing:
  - Can do this with very complex models. Discrepancies divide what we can reasonably model from what we care about.
  - Builds issues of approximate inference into the check. (The approximate posterior can be thought of as a truncated model.)
  - Treats model building as a first-class activity. We don't pretend that we stab in the dark for models and then pick the best one that we happened to stab.
  - Easy to implement. (Can be difficult to think about.)

- Some criticisms:
  - Uses the data twice. (A: Use held-out data. We have enough.)
  - The p-values are not calibrated. (A: Use graphs.)

  - Can't make hard decisions, e.g., rocket launches or drug approval. (A: Don't.)

- ML has given us complex models; this is a way to diagnose them.

- Unfortunately, ML focuses on things like predictive accuracy. Users of ML often care about exploration. PPCs—suitably made more modern—can bridge this gap.

- PUT THESE IN YOUR PROJECT: BIG BONUS.

6 Course summary

This is one strategy for data science / machine learning / statistics. (It's not the only one.)

(Put subjects from the class in our picture.)

We have emphasized

- Modularity—putting pieces together in more complicated models
- Generality—general-purpose inference, general modeling ideas.
- Scalability—online inference.
- Thinking about each data analysis problem individually—no cookbooks.