Exercises 4: Backfitting, Gibbs sampling, Hierarchical Models

Additive models and backfitting

Consider a multiple-regression problem with outcomes y_i and predictors $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$. An *additive model* takes the form

$$y = \alpha + f_1(x_1) + f_2(x_2) + \cdots + f_p(x_p) + \epsilon$$

for general functions f_j, where $E(\epsilon) = 0$ and $\text{var}(\epsilon) = \sigma^2$. Each individual effect can be nonlinear. But just as in linear regression, the effects from each predictor still add together to give the joint effect. The f_k are sometimes called partial response functions.

Suppose that someone hands you a set of good estimates for all f_j, $j \neq k$. Define the kth partial residual as the vector $y^{(k)}$ having elements

$$y_i^{(k)} = y_i - \alpha - \sum_{j \neq k} f_j(x_{ij}) \, .$$

Then we can clearly get a decent estimate for f_k by fitting $y^{(k)}$ versus x_k using the tools already in our kit (e.g. local linear regression).

To fit an additive model by *backfitting*, begin with an initial guess for all the f_j's. Successively refine your estimate for each partial response function f_k by computing the partial residuals $y^{(k)}$, and regressing these on x_k. Stop when the estimates reach some convergence criterion.[1]

The data in `air.csv` contain daily readings on some air-quality measurements for the New York area.[2]

Ozone: Average atmospheric ozone concentration in parts per billion from 1 PM to 3 PM at Roosevelt Island.

Solar.R: Solar radiation in Langleys in the wavelength range 400–770 nanometers from 8 AM to 12 PM at Central Park.

Wind: Average wind speed in miles per hour between 7 AM and 10 AM at LaGuardia Airport

Temp: Maximum daily temperature in degrees F at La Guardia Airport.

Write an R function to fit an additive model, and use it to regress ozone concentration on the other three variables.

At least two issues will need your attention:

1. If you subtract a constant c from f_j, and add that same constant to some other f_k, you will get the same regression function for all values of x. You will therefore need some way to identify them.

2. Should all dimensions have the same smoothing parameter?

[1] It is not obvious, at least to me, that this process converges to a unique solution when the predictor variables are correlated. But it does, for essentially the same reason (and under the same kinds of conditions) that the Gauss–Seidel algorithm works for solving linear systems.

[2] You'll notice that there are some missing days in there; we'll assume that these are missing at random.

Gibbs sampling

Consider a Bayesian analysis of the multiple linear-regression model, where $y = X\beta + \epsilon$. Suppose that the errors are assumed to be i.i.d. Gaussian, $\epsilon \sim \mathrm{N}(0, \sigma^2 I)$. Suppose that we specify an inverse-gamma prior for σ^2, and a simple hierarchical model for the regression coefficients:

$$\begin{aligned} (\beta \mid \tau^2) &\sim \mathrm{N}(0, \tau^2 I) \\ \sigma^2 &\sim IG(a/2, b/2) \\ \tau^2 &\sim IG(c/2, d/2) \end{aligned}$$

for fixed choices of a, b, c, d. Remember that an inverse-gamma prior for a variance v means that $1/v$ has a Gamma prior. It is most convenient to parametrize the Gamma distribution in terms of its shape and rate (not the scale). Thus if $1/v = r \sim \mathrm{Ga}(a, b)$, then $p(r) \propto r^{a-1} e^{-br}$.

(A) Derive the conditional posterior distributions for each model parameter: $p(\beta \mid y, \sigma^2, \tau^2)$; $p(\sigma^2 \mid y, \beta, \tau^2)$; and $p(\tau^2 \mid y, \beta, \sigma^2)$. Note that $p(\beta \mid y, \sigma^2, \tau^2)$ is actually a conditional distribution for the entire block of regression coefficients, rather than each coefficient individually.

(B) *Gibbs sampling*[3] is like a Bayesian version of backfitting: iteratively take a random draw from each parameter's conditional distribution, given the current values of all other parameters. Of course, unlike in backfitting, the draws will never converge to specific values as you run the algorithm for more iterations. Rather, they will build up a Monte Carlo sample from the joint posterior distribution over all parameters.[4]

Load the `diabetes` data set in the `BayesBridge` R package, available from CRAN. This is stored as a list, so to extract the responses and design matrix, you can use commands such as

```
Xd = diabetes$x
yd = diabetes$y
```

The outcome variable is a serum-insulin measurement in diabetes patients. The predictors are the patient's age, sex, BMI, and various other blood measurements. The x matrix has been standardized to have zero mean and unit ℓ^2 norm in each column. Fit a Bayesian linear model via Gibbs sampling to serum insulin versus the other predictors. Start with default values for the hyperparameters on σ^2 and τ^2 of $a, b, c, d = 1$. Remember to center the outcome, or to include a column in your design matrix for an intercept term.

[3] After Josiah Willard Gibbs, the father of modern thermodynamics. Why it is so named is a story for another day.

[4] This is even less obvious than the fact that backfitting converges. Formally, this process defines a Markov chain whose state space is the parameter space, and whose stationary distribution is (under suitable regularity conditions) the joint posterior. Gibbs sampling is a special case of Markov-chain Monte Carlo methods. A nice reference is *Monte Carlo Statistical Methods*, by Robert and Casella. An even better one is Peter Müller's course here at UT on Monte Carlo methods.

The following two papers might be interesting if you want some more background on choosing priors for variances in hierarchical models: (1) "Prior distributions for variance parameters in hierarchical models," by Gelman (Bayesian Analysis, 2006); and (2) "On the half-Cauchy prior for a global scale parameter," by Polson and Scott (Bayesian Analysis, 2012). Start with the first paper and only bother with the second if you really want to dig deeper here. These should be easy to find on the web.

Hierarchical models and shrinkage

Math tests

The data set in "mathtest.csv" shows the scores on a standardized math test from a sample of 10th-grade students at 100 different U.S. urban high schools, all having enrollment of at least 400 10th-grade students. (A lot of educational research involves "survey tests" of this sort, with tests administered to all students being the rare exception.)

Let θ_i be the underlying mean test score for school i, and let y_{ij} be the score for the jth student in school i. Starting with the "mathtest.R" script, you'll notice that the extreme school-level averages $\bar{y}_i$ (both high and low) tend to be at schools where fewer students were sampled.

1. Explain briefly why this would be.

2. Fit a normal hierarchical model to these data via Gibbs sampling:

$$\begin{aligned} y_{ij} &\sim \mathrm{N}(\theta_i, \sigma^2) \\ \theta_i &\sim \mathrm{N}(\mu, \tau^2) \end{aligned}$$

   Decide upon sensible priors for the unknown model parameters (μ, σ^2, τ^2).

3. Suppose you use the posterior mean $\hat{\theta}_i$ from the above model to estimate each school-level mean θ_i. Define the shrinkage coefficient κ_i as

$$\kappa_i = \frac{\bar{y}_i - \hat{\theta}_i}{\bar{y}_i},$$

   which tells you how much the posterior mean shrinks the observed sample mean. Plot this shrinkage coefficient for each school as a function of that school's sample size, and comment.

Price elasticity of demand

The data in "cheese.csv" are about sales volume, price, and advertisting display activity for packages of Borden sliced "cheese." The data are taken from Rossi, Allenby, and McCulloch's textbook on *Bayesian Statistics and Marketing.* For each of 88 stores (store) in different US cities, we have repeated observations of the weekly sales volume (vol, in terms of packages sold), unit price (price), and whether the product was advertised with an in-store display during that week (disp = 1 for display).

Your goal is to estimate, on a store-by-store basis, the effect of display ads on the demand curve for cheese. A standard form of a demand

curve in economics is of the form $Q = \alpha P^{\beta}$, where Q is quantity demanded (i.e. sales volume), P is price, and α and β are parameters to be estimated. You'll notice that this is linear on a log-log scale,

$$\log P = \log \alpha + \beta \log Q$$

which you should assume at least initially. Economists would refer to β as the price elasticity of demand (PED). Notice that on a log-log scale, the errors enter multiplicatively.

There are several things for you to consider in analyzing this data set.

1. The demand curve might shift (different α) and also change shape (different β) depending on whether there is a display ad or not in the store.
2. Different stores will have very different typical volumes, and your model should account for this.
3. Do different stores have different PEDs? If so, do you really want to estimate a separate, unrelated β for each store?
4. If there is an effect on the demand curve due to showing a display ad, does this effect differ store by store, or does it look relatively stable across stores?
5. Once you build the best model you can using the log-log specification, do see you any evidence of major model mis-fit?

Propose an appropriate hierarchical model that allows you to address these issues, and use Gibbs sampling to fit your model.

Gene expression over time

In `droslong.csv`, you will find a small subset of a time-course DNA microarray experiment. The gene-expression profiles of 2000 different genes in the fruit fly (Drosophila) genome are tracked over time during embryogenesis; you are getting data on 14 of these genes, organized in three groups (think of these as marking which cellular pathway that gene influences). For each gene at each time point, there are 3 "technical replicates"—that is, three copies of the same biological material from the same fly, run through the same process to measure gene expression.

The question of interest is: how does each gene's expression profile change over time, as the process of embryogenesis unfolds? Propose a hierarchical model for this data that properly reflects its structure. Fit this model using Gibbs sampling.

A nice graphics package is the "lattice" library. Install and load this; then try commands such as

```
xyplot(log2exp~time | gene, data=droslong)
```

```
xyplot(log2exp~time | group, data=droslong)
```

to begin exploring the structure of this data.

Data augmentation

Read the following paper:

> "Bayesian Analysis of Binary and Polychotomous Response Data." James H. Albert and Siddhartha Chib. *Journal of the American Statistical Association*, Vol. 88, No. 422 (Jun., 1993), pp. 669-679

The surefire way to get this paper is via access to JStor through the UT Library website. Let me know if this is an issue for you.

The paper describes a Bayesian treatment of probit regression (similar to logistic regression) using the trick of *data augmentation*—that is, introducing "latent variables" that turn a hard problem into a much easier one. Briefly summarize your understanding of the key trick proposed by this paper. Then see you if you can apply the trick in the following context, which is more complex than ordinary probit regression.

In "polls.csv" you will find the results of several political polls from the 1988 U.S. presidential election. The outcome of interest is whether someone plans to vote for George Bush (senior, not junior). There are several potentially relevant demographic predictors here, including the respondent's state of residence. The goal is to understand how these relate to the probability that someone will support Bush in the election. You can imagine this information would help a great deal in poll reweighting and aggregation (ala Nate Silver).

Use Gibbs sampling, together with the Albert and Chib trick, to fit a hierarchical probit model of the following form:

$$\begin{aligned} \Pr(y_{ij} = 1) &= \Phi(z_{ij}) \\ z_{ij} &= \mu_i + x_{ij}^T \beta \,. \end{aligned}$$

Here y_{ij} is the response (Bush=1, other=0) for respondent j in state i; $\Phi(\cdot)$ is the probit link function, i.e. the CDF of the standard normal distribution; μ_i is a state-level intercept term; x_{ij} is a vector of respondent-level demographic predictors; and β is a vector of state-invariant regression coefficients.

Note: there are severe imbalances among the states in terms of numbers of survey respondents! Following the last problem, the key is to impose a hierarchical prior on the state-level intercepts.