Hierarchical Models

David M. Blei

October 17, 2011

1 Introduction

- We have gone into detail about how to compute posterior distributions.

- Now we are going to start to talk about modeling tools—the kinds of components that can be used in data models on which we might want to compute a posterior.

- Notice that we can talk about models independently of inference.

- In this segment of the class, we'll introduce and review the other basic building blocks of models
  - Mixture models
  - Mixed membership models
  - Regression models
  - Matrix factorization models
  - Models based on time and space

2 What is a hierarchical model?

- There isn't a single authorative definition of a hierarchical model.

- Gelman et al., 2004
  - "Estimating the population distribution of unonobserved parameters"
  - "Multiple parameters related by the structure of the problem"

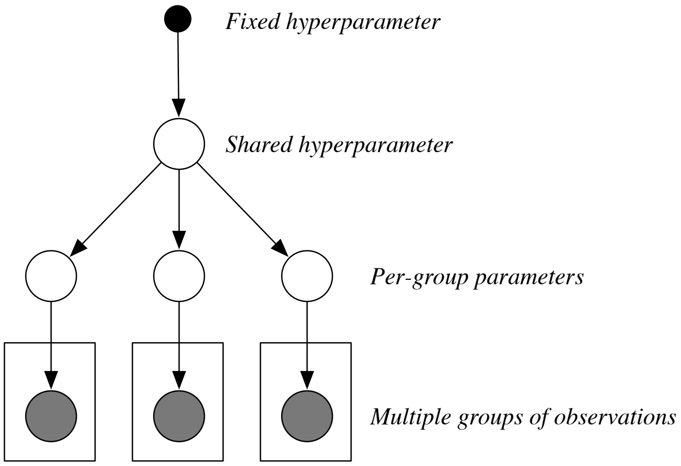
- Knowing something about one “experiment” tells us something about another.
  - Multiple similar experiments
  - Similar measurements from different locations
  - Several tasks to perform on the same set of images

- “Sharing statistical strength.” The idea here is that something we can infer well in one group of data can help us with something we cannot infer well in another. For example, we may have a lot of data from California but much less data from Oregon. What we learn from California should help us learn in Oregon.

- Key idea: Inference about one unobserved quantity affects inference about another unobserved quantity.
  - Includes some traditional hierarchical models
  - Does *not* include calling a prior/likelihood a hierarchical model
  - Includes models not necessarily thought of as hierarchical, such as HMMs, Kalman filters, mixtures of Gaussians.
  - (As such, it might be too forgiving a “definition.”)

3 The classical hierarchical model

- The classical hierarchical model looks like this:



- We observe multiple groups of observations, each from its own parameter.

- The distribution of the parameters is shared by all the groups. It too has a distribution.

- *Inference about one group's parameter affects inference about another group's parameter.*
  - You can see this from the Bayes ball algorithm.
  - Look back at the properties of hierarchical models. See how they apply here.
  - Consider if the shared hyperparameter were fixed. This is *not* a hierarchical model.

- Intuitive example: Height measurements of 8 year old daughters from different families.
  - Which family you are in will affect your height.
  - Assume each measurement is from a Gamma with an unknown per-family mean.
  - When we observe 1000 kids in a family, what do we know about the mean?
  - How about when we observe 10 kids in a family? How about 1?
  - What do we know about a new family that we haven't met?

- Let's show mathematically how information is transmitted in the predictive distribution.

- A more concrete (but still abstract) model
  - Assume m groups and each group has n_i observations.
  - Assume fixed hyperparameters α.

- Consider the following generative process—
  - Draw $\theta \sim F_0(\alpha)$.
  - For each group $i \in \{1,\ldots,m\}$:
    - Draw $\mu_i \,|\, \theta \sim F_1(\theta)$.
    - For each data point $j \in \{1,\ldots,n_j\}$:
      - Draw $x_{ij} \,|\, \mu_i \sim F_2(\mu_j)$.

- To make this more concrete, consider all distributions to be Gaussian, and the data to be real valued. Or, set F_0,F_1 and F_1,F_2 to be a chain of conjugate pairs of distributions.

- Let's consider the posterior distribution of the ith groups parameter given the data (including the data in the ith group).

  - Notice: we are not yet worrying about computation. We can still investigate properties of the model conditioned on data.

- The posterior distribution is

$$p(\mu_i \,|\, \mathscr{D}) \propto \int p(\theta \,|\, \alpha) p(\mu_i \,|\, \theta) p(\mathbf{x}_i \,|\, \mu_i) \left(\prod_{j \neq i} \int p(\mu_j \,|\, \theta) p(\mathbf{x}_j \,|\, \mu_j) d\mu_j \right) d\theta \tag{1}$$

- Inside the integral, the first and third term together are $p(\mathbf{x}_{-i}, \theta \,|\, \alpha)$. This can be expressed as

$$p(\mathbf{x}_{-i}, \theta \,|\, \alpha) = p(\mathbf{x}_{-i} \,|\, \alpha) p(\theta \,|\, \mathbf{x}_{-i}), \alpha). \tag{2}$$

  Note that $p(\mathbf{x}_{-i} \,|\, \alpha)$ does not depend on θ or μ_i.

- This reveals that the per-group posterior is

$$p(\mu_i \,|\, \mathscr{D}) \propto \int p(\theta \,|\, \mathbf{x}_{-i}, \alpha) p(\mu_i \,|\, \theta) p(\mathbf{x}_i \,|\, \mu_i) d\theta \tag{3}$$

- In other words, the other data influence our inference about the parameter for group i through their effect on the distribution of the hyperparameter.

- When the hyperparameter is fixed, the other groups do *not* influence our inference about the parameter for group i.

- Notice that in contemplating this posterior, the group in question does *not* influence its idea of the hyperparameter.

- Let's turn to *how* the other data influence θ. From our knowledge of graphical models, we know that it is through their parameters μ_j. However, we can also see it directly by unpacking the posterior $p(\theta \,|\, \mathbf{x}_{-i})$.

- The posterior of the hyperparameter θ given all but the ith group is

$$p(\theta \,|\, \mathbf{x}_{-i}, \alpha) \propto p(\theta \,|\, \alpha) \prod_{j \neq i} \int p(\mu_j \,|\, \theta) p(\mathbf{x}_j \,|\, \mu_j) d\mu_j \tag{4}$$

- (This gets fuzzy here.) Construe each integral as a weighted average of $p(\mu_j \,|\, \theta)$ weighted by $p(\mathbf{x}_j \,|\, \mu_j)$ for each value of μ_j.

- Suppose this is dominated by one of the values, call it $\hat{\mu}_j$.

- Then we can see the posterior as

$$p(\theta \,|\, \mathbf{x}_{-i}, \alpha) \propto p(\theta \,|\, \alpha) \prod_{j \neq i} p(\hat{\mu}_j \,|\, \theta) \tag{5}$$

- This looks a lot like a prior/likelihood set-up. E.g., if $p(\theta \,|\, \alpha)$ and $p(\hat{\mu}_j \,|\, \theta)$ form a conjugate pair then this is a conjugate posterior distribution of θ.

- It suggests that the hyperparameter is influened by the groups of data through what each one says about its respective parameter.

- This is not "real" inference, but it illustrates how information is transmitted.
  - Other groups tell us something about their parameters.
  - Those parameters tell us something about the hyperparameter.
  - The hyperparameter tells us something about the group in question.

- In Normal-Normal models, exact inference is possible.
  - However, in general $p(\alpha \,|\, \mathscr{D})$ does not have a nice form.
  - Notice that the GM is always a tree.
  - The problem is the functional form of the relationships between nodes.

- You can imagine how the variational inference algorithm transmits information back and forth. (Gibbs sampling and methods for exact inference do the same.)

- Consider other computations, conditioned on data.
  - The predictive distribution for a known group $p(x_i^{\text{new}} \,|\, \mathscr{D})$ will depend on other groups (for the hyperparameter) and on other data within the same group (for the per-group parameter).
  - The distribution of a parameter for a totally unobserved group μ^{new} will depend on the posterior of hyperparameter θ conditioned on all the data.

4 Hierarchical regression

- One of the main application areas of hierarchical modeling is to regression.

- Hierarchical (or multilevel) modeling allows us to use regression on complex data sets.
  - Grouped regression problems (i.e., nested structures)
  - Overlapping grouped problems (i.e., non-nested structures)
  - Problems with per-group coefficients
  - Random effects models (more on that later)

- Example: Collaborative filtering
  - Echonest.net has massive music data, attributes about millions of songs.
  - Imagine taking a data set of a user's likes and dislikes
  - Can you predict what other songs he/she will like or dislike?
  - This is the general problem of *collaborative filtering*.
  - Note: We also have information about each artist.
  - Note: There are millions of users.

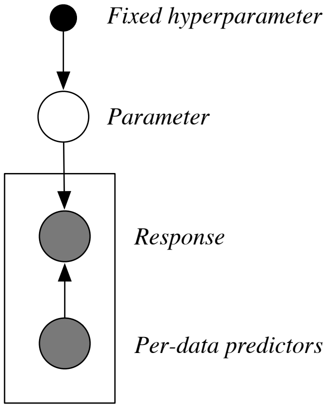
- Example: Social data
  - We measured the literacy rate at a number of schools.
  - Each school has (possibly) relevant attributes.
  - What affects literacy? Which programs are effective?
  - Note: Schools are divided into states. Each state has its own educational system with state-wide policies.
  - Note: Schools can be divided by their educational philosophy

- These kinds of problems—predictive and descriptive—can be addressed with regression.

- The causal questions in the descriptive analysis are difficult to answer definitively. However, we can (with caveats) still interpret regression parameters. (That said, we won't talk about causation in this class. It's a can of worms.)

4.1 Review

- Bayesian regression:



- Regression is a field. There are sequences of classes taught about it.

- Notation and terminology
  - *Response variable* y_i is what we try to predict.
  - *Predictor vector* x_i are attributes of the ith data point.
  - *Coefficients* β are regression parameters.

- The two regression models everyone has heard of are
  - Linear regression for continuous responses,
  $$y_i \mid x_i \sim \mathcal{N}(\beta^\top x_i, \sigma^2) \tag{6}$$
  - Logistic regression for binary responses (e.g., spam classification),
  $$p(y_i = 1 \mid x_i) = \text{logit}(\beta^\top x_i) \tag{7}$$
  - In both cases, the distribution of the response is governed by the linear combination of coefficients (top level) and predictors (specific for the ith data point).

- We can use regression for prediction.
  - The coefficient component β_i tells us how the expected value of the response changes as a function of each attribute.
  - For example, how does the literacy rate change as a function of the dollars spent on library books? Or, as a function of the teacher/student ratio?
  - How does the probability of my liking a song change based on its tempo?

- We can also use regression for description.

  - Given a new song, what is the expected probability that I will like it?
  - Given a new school, what is the expected literacy rate?

- Linear and logistic regression are examples of *generalized linear models*.
  - The response y_i is drawn from an exponential family,

$$p(y_i \mid x_i) = \exp\{\eta(\beta, x_i)^\top t(y_i) - a(\eta(\beta, x_i))\}. \tag{8}$$

  - The natural parameter is a function $\eta(\beta, x_i) = f(\beta^\top x_i)$.
  - In many cases, $\eta(\beta, x_i) = \beta^\top x_i$.

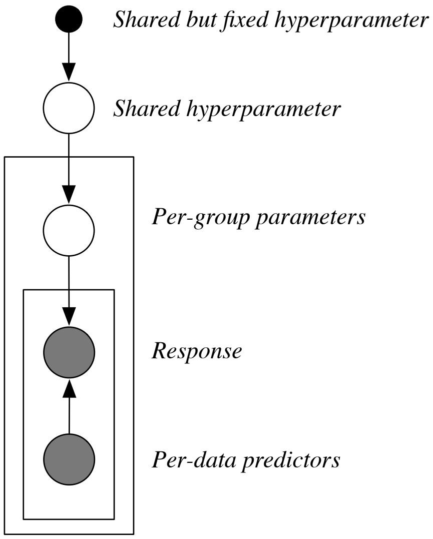
- Discussion
  - In linear regression y is Gaussian; in logistic regression y is Bernoulli.
  - GLMs can accommodate discrete/continuous, univariate/multivariate responses.
  - GLMs can accommodate arbitrary attributes.
  - (Recall the discussion of sparsity and GLMs in COS513.)

- Nelder and McCallaugh is an excellent reference about GLMs. (Nelder invented them.)

4.2 Hierarchical regression with nested data

- The simplest hierarchical regression model simply applies the classical hierarchical model of grouped data to regression coefficients.

- Nested regression:



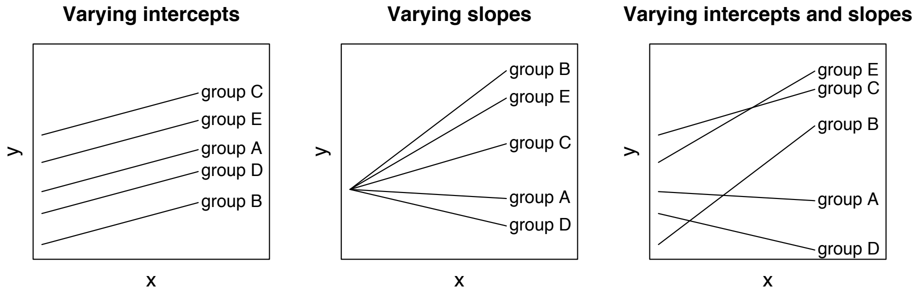
- Here each group (i.e., school or user) has its own coefficients, drawn from a hyperparameter that describes general trends in the coefficients.

- This allows for variance at the coefficient level.
  - Note that the variance too is interpretable.
  - An attribute can make a difference in different ways for different groups (high).
  - An attribute can be universal in how it predicts the response (low).

- But since β is multivariate we have some flexibility about which components to model hierarchically and which components to model in a fixed way for all the groups.

- One distinction we make in regression is between the "slope" term and "intercept" term. The intercept term is (effectively) for a coefficient that is always equal to one.

- Gelman and Hill (2006) draw the following useful picture:

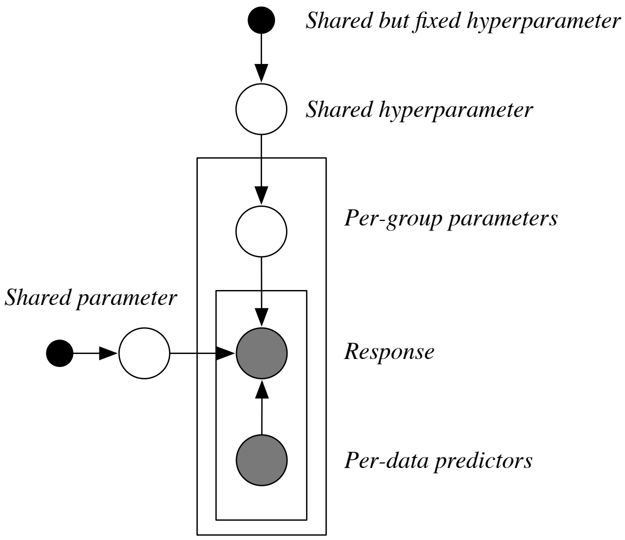


- This corresponds to modeling intercept, slope, or both coefficients hierarchically.

- The same choices apply for each cofficient or subset of coefficients. Often the problem will dictate where you want group-level coefficients or top-level coefficients. (Hierarchical regression means having to make more choices!)

- Also, the group level parameters can be the noise term, i.e., σ^2. Thus you can model errors that are correlated by group. (Thanks Matt S. for pointing this out.)

- More complicated nested regression:



- Back to the literacy example—
  - Responses are grouped by state.
  - Each state might have a different base literacy rate (intercept).
  - It increases or decreases in the same way as other schools, based on the attributes.
  - Or, the relationship between some attributes and literacy rates is the same for all states; for others it varies across state.

- Back to the collaborative filtering example—
  - Responses are grouped by artist.
  - Each artist has a different base probability that the user will like them.
  - It changes in the same way as other artists, based on attributes of the song.
  - Or, the relationship between some attributes and probability is common for all artists; for other attributes it varies by artists.

- Recall the results from the previous section, and how they apply to regression.
  - What I know about one school tells me about others.
  - What I know about one artist tells me about others.
  - I can make a prediction about a new artist, even if they have new coefficients.
  - I can make a prediction about a new school, before observing data from it.

- A note on computation

  - As above, Normal priors and linear regression gives itself to exact inference.
  - However, other response variables (e.g., binary) and other priors do not.
  - The normal prior can be limiting, especially if you want to use tools like Laplace priors or spike and slab priors to find sparse solutions. (Recall COS513.)

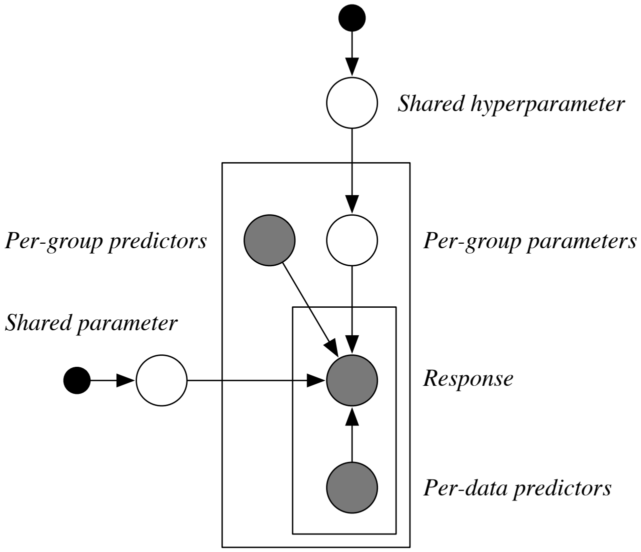
4.3 Nested data with per-group covariates

- Hierarchical regression immediately lets us reason about groups.

- In classical regression, per-group (e.g., state) attributes are encoded within the data. For example, how many dollars does this state apply to literacy programs?

- Encoding them in the groups themselves is more natural.

- The most complicated nested regression I can think of:



  The shared parameter involves the per-group predictors (and possibly the per-data predictors). The per-group parameters involves some or all of the per-data predictors.

- Also consider artists in the collaborative filtering example.
  - We might have descriptions of each artist (in fact, we do).
  - We can form a term vector and predict the “baseline” of whether the user will like a song by that artist. This helps predictions on new songs by new artists.

4.4 Regression with non-nested data

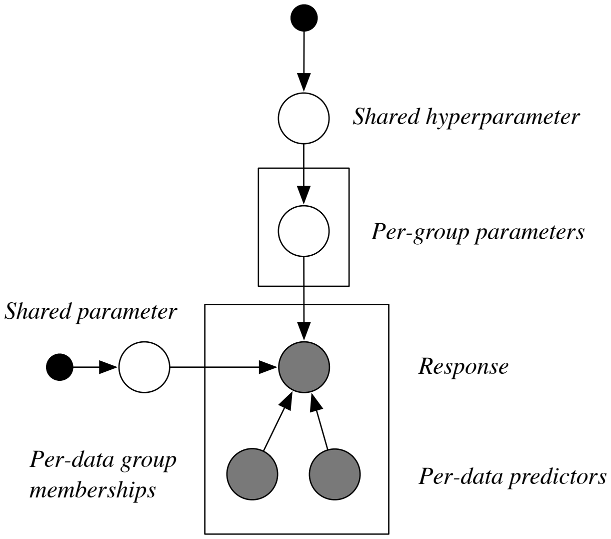
- So far, we considered *grouped* or *nested* data.

- Hierarchical modeling can capture more complicated relationships between the hidden variables. Thus, it can capture more complicated relationships between the observations.

- Suppose each data point is associated with multiple groups.
  - The per-data coefficients are a sum of their group coefficients.
  - Suppose there are G possible groups.
  - Let g_i be the per-data group memberships, a G binary vector. These memberships are given, along with the per-data predictors x_i.
  - In this model,

$$\begin{aligned} \beta_g &\sim F_0(\alpha) \quad \text{for } g \in \{1, \ldots, G\} && (9) \\ y_i &\sim F_1((g_i^\top \beta)^\top x_i) && (10) \end{aligned}$$

- Graphical models are less useful here—the structure of how the data relate to each other is encoded in g_i.

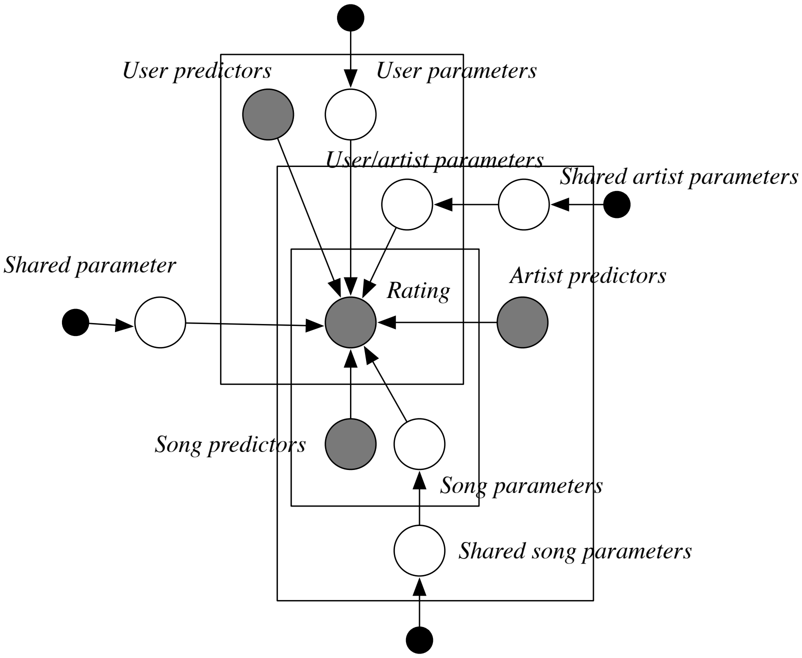


- Here we don't *need* the per-group parameters to come from a random hyperparameter. They exhibit sharing automatically because data are members of multiple groups.

- And it can get more complicated still:

- Local parameters in a time series
- Coefficients that aren't observed (i.e., random effects)
- Parameters in a tree
- Hierarchies in hierarchies

- For example, a mammoth model of the Echonest data



  - Multiple users allow many kinds of parameters and predictors
  - Songs are within artists; users rate all songs
  - Some songs might be universally popular
  - Some artists might be popular within users
  - Anything is possible
  - Searching within models is difficult...

- But first, let's get back to a simple hierarchical model from the 1950s.

5 Empirical Bayes and James-Stein estimation

- An empirical Bayes estimate is a data-based point estimate of a hyperparameter.

- Gelman et al. point out that this is an approximation to a hierarchical model, i.e., when we have a posterior distribution of a hyperparameter.

- But empirical Bayes has an interesting history—it came out of frequentist thinking about Bayesian methods. And, empirical Bayes estimators have good frequentist properties.

- Brad Efron's 2010 book *Large-Scale Inference* begins as follows: "Charles Stein shocked the statistical world in 1955 with his proof that maximum likelihood estimation methods for Gaussian models, in common use for more than a century, were inadmissable beyond simple one- or two-dimensional situations."

- An *admissable* estimator has the property that no other estimator has risk as small.

- Specifically, an estimator $\hat{\theta}$ is inadmissable if there exists another estimator $\hat{\theta}'$ such that

$$\begin{aligned} R(\theta, \hat{\theta}') &\leq R(\theta, \hat{\theta}) \quad \text{for all } \theta \\ R(\theta, \hat{\theta}') &< R(\theta, \hat{\theta}) \quad \text{for at least one } \theta \end{aligned}$$

- We follow Efron's explanation of James-Stein estimation as an empirical Bayes method. (In many places here, we mimic his wording too. Efron is as clear as can be.)

- Consider the following hierarchical model

$$\begin{aligned} \mu_i &\sim \mathcal{N}(0, A) && (11) \\ x_i \,|\, \mu_i &\sim \mathcal{N}(\mu_i, 1) \quad i \in \{1, \ldots, n\}. && (12) \end{aligned}$$

- Note that the data are not grouped. Each comes from an individual mean.

- The posterior distribution of each mean is

$$\mu_i \,|\, x_i \sim \mathcal{N}(B x_i, B), \quad \text{where } B = A/(A+1). \tag{13}$$

- Now, we observe $x_{1:n}$ and want to estimate $\mu_{1:n}$. We estimate $\hat{\mu}_i = f_i(x_{1:n})$.

- Use total squared error loss to measure the error of estimating each μ_i with each $\hat{\mu}_i$,

$$L(\mu_{1:n}, \hat{\mu}_{1:n}) = \sum_{i=1}^{n} (\hat{\mu}_i - \mu_i)^2 \tag{14}$$

- The corresponding risk—risk is the expected loss under the truth—is

$$R(\mu_{1:n}) = \mathrm{E}[L(\mu_{1:n}, \hat{\mu}_{1:n})] \tag{15}$$

  where the expectation is taken with respect to the distribution of $x_{1:n}$ given the (fixed) means $\mu_{1:n}$. (Recall that the loss is a function of $\hat{\mu}_{1:n}$. That is a function of the data. And, the data are random.)

- The maximum likelihood estimator simply sets each $\hat{\mu}_i$ equal to x_i,

$$\hat{\mu}_i^{\text{MLE}} = x_i. \tag{16}$$

- In this case, the risk is equal to the number of data,

$$R^{\text{MLE}}(\mu_{1:n}) = n. \tag{17}$$

  Notice this does not depend on $\mu_{1:n}$.

- Now, let's compare to the Bayesian solution.

- Suppose our prior is real, i.e., $\mu_i \sim \mathcal{N}(0, A)$. Then the Bayes estimate is

$$\hat{\mu}_i^{\text{Bayes}} = Bx_i = \left(1 - \frac{1}{A+1}\right) x_i \tag{18}$$

- For a fixed $\mu_{1:n}$, this has risk

$$R^{\text{Bayes}}(\mu_{1:n}) = (1-B)^2 \left(\sum_{i=1}^{n} \mu_i^2\right) + nB^2 \tag{19}$$

- The Bayes risk is the expectation of this with respect to the "true" model of $\mu_{1:n}$,

$$R^{\text{Bayes}} = n \left(\frac{A}{A+1}\right). \tag{20}$$

- Note: the Bayes risk of the MLE is simple $R^{\text{MLE}} = n$.

- So, if the prior is true then the benefit is

$$R^{\text{MLE}} - R^{\text{Bayes}} = n/(A+1) \tag{21}$$

  For example, Efron points out, if $A = 1$ then we reduce the risk by two.

- But, of course, we don't know the prior. Let's estimate it.

- Suppose the model is correct. The marginal distribution of x_i is

$$x_i \sim \mathcal{N}(0, A+1). \tag{22}$$

  We obtain this by integrating out the prior.

- The sum of squares $S = \sum_{i=1}^n x_i^2$ is a chi-square with n degrees of freedom

$$S \sim (A+1)\chi_n^2. \tag{23}$$

- This means that

$$\text{E}[(n-2)/S] = 1/(A+1) \tag{24}$$

- If we substitute $(n-2)/S$ for $1/(A+1)$ in the Bayes estimate, we obtain the James-Stein estimate

$$\hat{\mu}_i^{\text{JS}} = \left(1 - \frac{n-2}{S}\right) x_i \tag{25}$$

- The risk of this estimator is

$$R_A^{\text{JS}} = nA/(A+1) + 2/(A+1) \tag{26}$$

  This is bigger than the Bayes risk, but still smaller than the MLE risk for big enough sample sizes. Note that

$$R_A^{\text{JS}} / R_A^{\text{Bayes}} = 1 + 2/(nA) \tag{27}$$

- This defines James-Stein but still assumes a model where the means are centered at zero. The result that "shocked the statistical world" is that for $n \geq 3$,

$$\text{E}\left[\sum_{i=1}^n (\hat{\mu}_i^{\text{JS}} - \mu)^2\right] < \text{E}\left[\sum_{i=1}^n (\hat{\mu}_i^{\text{MLE}} - \mu^2\right], \tag{28}$$

  for every choise of $\mu_{1:n}$ and where expectations are taken with respect to n data points drawn from $\mu_{1:n}$.

- Note that this result holds without regard for your prior beliefs about $\mu_{1:n}$.

- Keep reading Efron—
  - A version holds when $n \geq 4$ where the mean is estimated as well. This leads to a version for regression.
  - It works *on average*, but can really misestimate individual data.
  - This is why it didn't supplant the MLE.
  - But this is why it is important now in large-scale data settings.