A Method for Assigning Letter Grades: Multi-Curve Grading

Alex Strashny*
University of California, Irvine

May 2, 2003

Abstract

This paper describes a new method for assigning letter grades to students based on their raw scores, which I call Multi-Curve Grading (MCG). The intuition behind the method is that a class can be composed of several different subgroups, each of which should be assigned a different grade. In this, the method quantifies and builds upon the Distribution Gap grading method.

I model the raw scores as coming from a Normal Mixture, with each component of the mixture corresponding to a different letter grade. I estimate this model using Gibbs Sampler. Based on this model, I calculate the probability that each student's raw score corresponds to each possible letter grade. The grader's degree of leniency is used to specify his loss function and thus to assign the most optimal letter grades.

I compare Multi-Curve Grading to other common grading methods, such as the Standard Deviation method. It appears to assign grades better than these other methods.

Keywords: letter grade, course grade, grade assignment, education measurement, mixture model, Gibbs Sampler, Markov Chain Monte Carlo
JEL: I29, C11, C51

*Dept. of Economics, University of California - Irvine, 3151 SSPA, Irvine, CA 92697-5100. [astrashn@uci.edu]. Thanks to Justin Tobias and William Batchelder for their helpful comments.

1 Introduction

At the end of a course, the grader has to assign a letter grade to each student in the class based on the student's performance. This performance in the course is often summarized by one or several *raw scores*. For example, these could be the scores from a midterm and a final exam. The grader then has to use these raw scores to meaningfully assign a *letter grade*, such as A, or $B-$, to each of student. This paper addresses this problem of assignment assuming that only the raw scores, and no other information, are available.

While there is no single, universally agreed upon, grade assignment method, several methods are very common. Among these are "grading on a curve" and Adjusted Scale. Some graders see what grade a couple of different methods would assign, and then assign the grade that is highest among these. This procedure reveals that these graders would rather overestimate the assigned grade than underestimate it.

Some more experienced graders assign letter grades by "staring" at the raw scores. A natural way for doing this is to make grade cutoffs in places where few students score. This practice is called the Distribution Gap method. The problem with it is that it is highly subjective and often unusable in practice. It does, however, point to the possibility that a class is composed of several subgroups of students. I use this idea in *Multi-Curve Grading* (MCG).

In section 2, I briefly describe some grading methods commonly used today, as well as their advantages and disadvantages. In section 3, I describe a mixture model inspired by the Distribution Gap method. I explain how the model is estimated in appendix A.

In section 4, I use the estimated model to find the probability that a student should be assigned each possible letter grade. In section 5, I explain how to assign letter grades by minimizing the grader's loss function. The loss function that I consider allows the grader to reflect his degree of leniency.

In section 6, I test MCG with real data. I compare the letter grades assigned by MCG to those actually assigned by the grader. It turns out that MCG does better than the conventional methods considered.

2 Conventional grading methods

In this section, I briefly describe conventional grading methods, along with their advantages and disadvantages. For more details, see Frisbie & Waltman 1992. The method presented in this paper builds on the Distribution Gap Method.

Note that these descriptions assume that there is only one raw score

per student. If there are several raw scores, most graders first combine them into a single score and then use that score for assignment. This may not be strictly correct. See Cross 1995 and section 4.2 for details.

2.1 Straight Scale

If the raw score falls within a certain predetermined interval, assign the letter grade corresponding to that interval. See table 1 for a commonly used Straight Scale. The method is very easy to apply. However, the intervals are created before any actual raw scores are seen. That is, this method does not even consider the actual raw scores in making the cutoffs between letter grades.

Also, Straight Scale makes no sense if the test that produced the raw scores is either too hard or too easy. But often it's impossible to know the difficulty of a test until after the raw scores are seen.

2.2 Standard Deviation

A letter grade is given based on how far, in standard deviations, the student's raw score is from the mean of all raw scores. See table 1 for a commonly used Standard Deviation scale. The advantage of this method is that it automatically adjusts letter grades to the difficulty of the test that produced the raw scores. For example, if a test is made more difficult, and the only result is that the mean of raw scores decreases, all the letter grades would be the same.

However, there is some arbitrariness in the method: the mean is set to an arbitrary letter grade; the standard deviation distance between letter grades is also arbitrary.

Further, this method implicitly assumes that all the raw scores come from the same population. If the raw scores come from a single population, then the sample mean and variance are estimators of corresponding population moments. However, if the raw scores come from more than one subpopulation, these sample quantities are not estimators of any meaningful population quantities.

2.3 Curve

The Curve method is the same thing as the Standard Deviation method, with the further assumption that the population that all the raw scores come from is Normal. Given this assumption, we know at what percentiles to make grade cutoffs. See table 1. In addition to having all the disadvantages of the Standard Deviation method, the Curve method makes an additional assumption that may or may not be true. This method is not discussed further in this paper, as it is a special case of the Standard Deviation method.

2.4 Modifications and Combinations

Many graders, at a loss as to which method to use, try to either modify or combine several common methods. Below I describe three common adjustments. These adjustments can be much more complicated. For an example, see Birnbaum 2001.

2.4.1 Adjusted Scale

All raw scores are divided by the highest raw score and multiplied by 100. Letter grades are then assigned using Straight Scale based on these adjusted scores. This method addresses the problem of a test that produced raw scores that are "too low". However, it is unclear why the raw scores must be divided by the highest score and not, for instance, the second highest. The grades of all the students in the class are very sensitive to the performance of the best student.

2.4.2 Adjustments to Standard Deviation

Adjustments can be made to the Standard Deviation method. For instance, the mean could be set to B rather than C.

2.4.3 Maximum

Some graders give each student the highest grade that he could receive under a couple of different grading methods. For example, the students might receive the highest grade under either Adjusted Scale or Standard Deviation methods. This approach reveals that graders tend to be lenient.

2.5 Distribution Gap

Raw scores are plotted on a histogram. Grade cutoffs are made in places where few students have scored. Unlike the Straight Scale, this method assigns cutoffs after looking at the raw scores. Thus it doesn't matter if the test that produced the raw scores is too hard or too easy. Also, unlike the Standard Deviation method, Distribution Gap does not assume that there is a single underlying population.

The disadvantage of Distribution Gap is that it is hard to use in practice. Gaps appear and disappear depending on the size of histogram bins. Deciding at which gaps to make the cutoffs is subjective. Finally, there just might not be any obvious gaps.

Still, this method gives the valuable insight that the underlying population might be composed of several components. The method also states that each component should be assigned a different letter grade.

2.6 Staring

Many experienced graders don't strictly follow any mechanical method. Rather, they determine grade cutoffs based on "feel". Because feel varies from grader to grader, I cannot quantify this method. However, I can measure how closely grades assigned by various mechanical methods come to grades assigned by Staring in some specific examples. This is what I do in section 6.

3 Mixture model basis of Multi-Curve Grading

3.1 Model setup

In all, there are 13 possible letter grades. They are, from lowest to highest, $\{F, D-, D, D+, C-, C, C+, B-, B, B+, A-, A, A+\}$. The numeric equivalents of these grades are $\{0.0, 0.7, 1.0, 1.3, 1.7, 2.0, 2.3, 2.7, 3.0, 3.3, 3.7, 4.0, 4.3\}$.

I assume that the raw scores come from a Normal Mixture Model. That is, the distribution of scores is a weighted sum of several Normal distributions. Each component of the mixture corresponds to a different letter grade. Thus, since there are thirteen letter grades, I assume that the mixture has thirteen components. This assumption means that all thirteen letter grades are potentially present. This does not, however, mean that all the letter grades will actually be assigned. It is possible that in the end, a letter grade will not be assigned to any student.

Raw scores are distributed independently and identically with the distribution

$$p(x_i) = \sum_{g=1}^{G} \pi_g \phi(\mu_g, \sigma_g^2), \tag{1}$$

where x_i is the raw score of student i; g indexes the $G = 13$ components of the mixture; π_g is the component probability of component g, $\sum \pi = 1$; $\phi(\cdot)$ is the Normal pdf; μ_g and σ_g^2 are mean and variance of component g. In this application, a very natural restriction is that the components are ordered by their mean. That is,

$$\mu_1 < \mu_2 < \ldots < \mu_G. \tag{2}$$

This model is motivated by the observation, underlying the Distribution Gap method, that there are often gaps in scores. If the component means are relatively far from each other, then these gaps are apparent to the naked eye. Under Distribution Gap, cutoffs are made at the gaps. This is similar to assigning a different letter grade to each component.

There are three parameters associated with each component. However, one of the component probabilities π_g can always be determined if we know the other component probabilities. Thus, this model has a total of $3G - 1$ parameters.

I estimate the model using a technique called Gibbs Sampler. Details of the estimation procedure are given in appendix A. Obviously, the more raw scores there are, the more accurately the model is estimated. However, since the estimation procedure is Bayesian, the model can always be estimated, no matter how small the data set is.

Figure 1 shows a histogram of raw scores for a particular class of students, along with the estimated distribution. The lower lines indicate each component. The upper line is the combined distribution, the sum of all components. Note that the estimated distribution is nowhere near bell-shaped.

3.2 On the arbitrariness of thirteen components

The assumption that the mixture model has thirteen components might seem arbitrary. There are two approaches to justifying this assumption.

First, all models impose some view of the data. For example, the Standard Deviation method implicitly assumes that the data comes from a single population. It is impossible to have a model that does not influence our view of the data.

However, some models are more rigid while others are more accommodating to the data. A model that assumes a single population should not properly be used with data that in reality is composed of two or more subpopulations. But if the data truly comes from just one population, a model that allows for several subpopulations can still be used. In other words, the assumption made in this paper is less restrictive than the assumption underlying the popular Standard Deviation method.

Second, while having thirteen components in the mixture is arbitrary, the fact that there are thirteen possible letter grades is also arbitrary. Allowing thirteen letter grades implicitly makes the assumption that there can be up to thirteen subpopulations. Thus, having thirteen components simply *reflects* this existing assumption. It is not creating a new assumption.

4 Probability distribution of letter grade

4.1 Calculation

Here I explain how to calculate, based on the above model, the probabilities that each raw score corresponds to each possible letter grade. These probabilities are then used to assign the letter grade.

The probability that a particular raw score belongs to a component of the mixture is proportional to the ordinate of that component at that raw score. In other words,

$$\Pr(x_i \in g) \propto \pi_g \phi(x_i | \mu_g, \sigma_g^2). \tag{3}$$

For each raw score x_i, calculate the above probability for all $G = 13$ components. From this, calculate the probability that the raw score belongs to each of the components. Because the thirteen model components correspond to the thirteen letter grades, the probability that a raw score belongs to a particular letter grade is the same as the probability that it belongs to the corresponding component.

Example 1 *Refer to figure 1. Let's consider the raw score of* 60. *By visual inspection, we see that the sixth component, that is, the component corresponding to C, is the most probable. Table 2 lists the ordinates of each component at* 60. *For instance, the height of the sixth component at* 60 *is* 0.0058.

Next, the table lists probabilities that the raw score comes from each component. These probabilities are just the ordinates normalized so that their sum is equal to 100%. *Thus, the probability that* $x = 60$ *comes from the sixth component is* 38%.

4.2 Several raw scores per student

The preceding discussion describes how, for each student, to calculate the *probability distribution of letter grade* based on a single raw score. Often, each student has several raw scores coming from, for instance, several exams. The grader might want to give a particular weight to each raw score.

One way of dealing with this is to calculate, for each student, the weighted average of raw scores and then to just work with that. This is what graders commonly do, especially if they use the Staring method.

This approach, however, has been widely criticized. See, for instance, Cross 1995. The problem with this approach is that raw scores that come from an exam with lower variance have a lower influence on the letter grade.

To avoid this, calculate the probability distribution of letter grade for each raw score. Then, take the weighted average of the probability distributions. Thus, no matter how many raw scores per student there are, in the end, each student has only one probability distribution of letter grade.

Example 2 *Suppose a student took two tests. The grader wants to weigh the first test at* 40% *and the second at* 60%*. To do this, he first constructs the probability distribution of letter grade based on each test. He then combines these distributions using the weights. He adds the first distribution multiplied by* 40% *to the second distribution multiplied by* 60%*.*

5 Estimating letter grade

5.1 Loss function

A loss function describes the "loss" that the grader experiences if he assigns a certain letter grade while the student truly "deserves" another letter grade. An optimally assigned letter grade is one that minimizes expected loss based on the probability distribution of letter grade that is described above.

The loss function that is used in many applications is the quadratic loss. This function states that the loss is the square of the distance between the true parameter and its point estimate. Quadratic loss is *symmetric*. This means that underestimation produces the same loss as overestimation.

Quadratic loss may not be appropriate in the particular application of assigning letter grades. This is because a grader might feel worse if he assigns a grade that's too low than if he assigns a grade that's too high. After all, a low grade will influence the student's future adversely. This tendency towards leniency is reflected in the Maximum method for assigning grades.

Since the grader typically feels a different loss for overestimating and underestimating the letter grade, his loss function is, in general, *asymmetric*. Because of this, I use the following loss function:

$$C(\hat{y}_i, y_i) = \begin{cases} c|\hat{y}_i - y_i| \text{ if } \hat{y}_i \leq y_i \\ |\hat{y}_i - y_i| \text{ if } \hat{y}_i > y_i \end{cases}. \tag{4}$$

Here, y is the numeric equivalent of the letter grade that the student truly deserves; $\hat{y}$ is the numeric equivalent of the grade that the grader assigns; and c is a positive constant that reflects grader's preferences. When $c = 1$, the grader feels equally badly about overestimating and underestimating the grade. When $c > 1$, the grader feels worse about underestimating than he feels about overestimating.

Let $q = \frac{c}{c+1}$. The optimal letter grade under this loss function is the q-th quantile of probability distribution of letter grade. Since the distribution is not continuous, I choose the highest letter grade whose

cumulative probability is less than q. When $q = 0.5$, the loss is symmetric, and the optimal letter grade is the median. If $q > 0.5$, the grader feels a higher loss if he underestimates, and so he "bumps up" the grade. The situation when $q > 0.5$ can be interpreted as a higher leniency in grading. Because of this, I refer to q as the *Leniency Factor* (LF). Leniency Factor of 0.5 means that the grader is neither lenient nor strict. Leniency Factor between 0.0 and 0.5 means that the grader is strict. Leniency Factor between 0.5 and 1.0 means that the grader is lenient.

A grader can easily calculate his Leniency Factor based on how he feels about overestimating and underestimating the letter grade. In the formula for Leniency Factor, c is the loss that the grader feels for underestimating as compared to the loss he feels for overestimating. For example, if the grader cares about underestimating the grade one and a half times as much as he cares about overestimating it, then the Leniency Factor is $\frac{1.5}{1.5+1} = 0.6$.

For convenience, if a grader's LF is 0.5, I say that he is in neutral mode, while if it's 0.6, I say that he is in lenient mode.

Example 3 *Consider the probability distribution of letter grade along with the cumulative distribution in table 2. If the grader is in neutral mode, the most optimal letter grade is C−. If he is in lenient mode, the optimal grade is C. Assigning B would require an incredibly high Leniency Factor of* 0.998 *or above.*

5.2 Discussion

In MCG, unlike in the Standard Deviation method, the grader does not have to arbitrarily set the mean to a particular letter grade. Rather, he sets the Leniency Factor, which is a more intuitive quantity, as it directly describes the grader's own preferences. The grader only has to ask himself about his relative preference for overestimating and underestimating the grade. The answer to this question directly gives the Leniency Factor.

Suppose a grader wants to use the Standard Deviation method and he is considering setting the mean to a particular letter grade. He can use MCG to find what values of the Leniency Factor would make the mean equal to the letter grade that he is considering. He can then check if his assignment makes sense by seeing if the Leniency Factor truly reflects his preferences.

For example, suppose a grader is considering setting the mean to a B. Without seeing the actual raw scores, this might sound reasonable. However, for a particular distribution of raw scores, it might turn out that this assignment requires the Leniency Factor to be 0.9. This high

value implies that the grader cares about underestimation nine times more than he cares about overestimation. Such preferences are clearly unreasonable. Thus, MCG reveals that, in this particular case, setting the mean to a B is extremely lenient.

6 Evaluating MCG with data

In this part, I give two real life examples. I compare letter grade assignments from MCG to the grades actually assigned by graders. The reader can judge how well MCG does by visual inspection. I use two different measures to quantify how well each grading method does.

Above, I discuss optimal assignment of grades when raw scores come from more than one source. I mention that, strictly speaking, it is inappropriate to assign grades based on the combined raw score. However, in all the examples in this part, graders actually assigned grades based on the combined raw score. For comparison purposes, I consider how well MCG does also based on the combined score.

6.1 Measuring performance

There are two measures that seem "natural" for determining how well a grading method does. First, we can simply look at the average loss as defined in equation 4. Let *Class Loss* be

$$CC = \frac{1}{N}\sum_{i=1}^{N} C_i, \tag{5}$$

where N is the number of students in a class. I interpret the grades actually assigned by the grader as the "true" grades. This way, a lower Class Loss means that the grading method assigns grades closer to those actually assigned by the grader.

Another way of evaluating performance is by looking at the *raw coefficient of determination*, R_r^2. Let e_i be the difference between the numeric equivalents of the letter grade assigned by the grader and the letter grade assigned by a grading method. That is, $e_i = y_i - \hat{y}_i$. Then

$$R_r^2 = 1 - \frac{\sum_{i=1}^{N} e_i^2}{\sum_{i=1}^{N} y_i^2}. \tag{6}$$

R_r^2 is always between 0% and 100%. Higher values indicate that a grading method gets closer to the grades actually assigned.

Below, I compare MCG to some conventional grading methods in neutral (Leniency Factor 0.5) and lenient (LF 0.6) modes.

6.2 Case 1: Large Class

A certain class consists of 374 students. The mean raw score is 70; the median is 74.4; the standard deviation is 17.1. Figure 1 shows a histogram of the scores along with an estimated distribution. Note that the distribution is not bell-shaped.

Of the conventional methods considered, grade assignments made by the grader are closest to Standard Deviation with mean set to B. Table 3 shows cutoffs assigned by the grader and those assigned by MCG.

The mean letter grade assigned by MCG in neutral mode is 2.9, or slightly below B; in lenient mode, the mean is 3.0, or B. In neutral mode, 8.8% of the students receive a failing grade (below a $C-$). In lenient mode, 8.6% of the students fail. The grader fails 9.3% of the students.

Table 4 quantifies the performance of several different grading methods. In neutral mode, MCG has the lowest Class Loss (just 0.14). This is even lower than Class Loss under Standard Deviation with mean at B (0.16). In lenient mode, MCG is slightly worse, but is still much better than four out of the five conventional grading methods that are considered. R_r^2 tells the same story.

6.3 Case 2: Medium Class

Another class is made up of 129 students. The mean and median raw scores are around 64.7; the standard deviation is 6.9. Figure 2 shows a histogram of scores along with an estimated distribution.

Of the conventional methods considered, grade assignments made by the grader are closest to Standard Deviation with mean set to B. However, he did not assign any failing grades. Table 5 shows cutoffs assigned by the grader and those assigned by MCG.

The mean letter grade assigned by MCG in neutral mode is 2.8, or about $B-$; in lenient mode, the mean is 2.9, or just below B. Both in neutral and lenient modes, 2.3% of the students fail.

Table 6 quantifies the performance. MCG beats the other grading methods considered both in neutral and lenient modes.

7 Conclusion

In this paper, I develop a method for assigning letter grades based on raw scores. The method is inspired by an existing method called Distribution Gap. MCG basically removes the subjectivity from Distribution Gap, making it more applicable. MCG reflects the common belief that a class is composed of several sub-populations, each of which should be assigned a different letter grade.

To apply MCG, a grader needs only to determine his own Leniency Factor. This is an intuitive measure that reflects how leniently the grader wants to assign letter grades. A value of 0.5 is neutral. Values above this are lenient. If the grader is lenient, a value around 0.6 is suggested.

I bring a couple of cases in which an experienced grader already assigned letter grades by eye. I then use MCG and some conventional grading methods to assign letter grades based on the same raw scores. MCG gets very close to what the grader actually assigned, and performs extremely well as compared to other grading methods. The advantage of MCG is that a grader who is using it does not have to be experienced and does not have to spend any time going through all the raw scores – all the work of assignment is done automatically by a computer.

In short, I believe MCG to be sound, as it's founded on other methods; it is easy to use, as all the work is done by a computer; and it assigns grades accurately.

References

Birnbaum, D. J. 2001, 'Grading system for Russian fairy tales'.
URL: *http://clover.slavic.pitt.edu/~tales/02-1/grading.html*

Carlin, B. P. & Louis, T. A. 2000, *Bayes and Empirical Bayes Methods for Data Analysis*, second edn, Chapman and Hall.

Cross, L. H. 1995, *Grading Students*, ED398239, ERIC Clearinghouse on Assessment and Evaluation, Washington, DC.
URL: *http://www.ericfacility.net/ericdigests/ed398239.html*

Division of Measurement and Evaluation, University of Illinois 2003, *Assigning Course Grades*.
URL: *http://www.oir.uiuc.edu/dme/Exams/ACG.html*

Frisbie, D. A. & Waltman, K. K. 1992, 'Developing a personal grading plan', *Educational Measurement: Issues and Practice* .
URL: *http://depts.washington.edu/grading/plan/frisbie1.htm*

Raftery, A. E. 1996, Hypothesis testing and model selection, *in* S. R. W.R. Gilks & D. Spiegelhalter, eds, 'Markov Chain Monte Carlo in Practice', Chapman and Hall, pp. 163–188.

Spencer, C. 1983, 'Grading on the curve', *Antic* **1**(6), 64.
URL: *http://www.atarimagazines.com/v1n6/education.html*

Stephens, M. 1997, Bayesian Methods for Mixtures of Normal Distributions, PhD thesis, Magdalen College, Oxford.

A Estimating mixture model parameters

Here, I explain how to estimate the model given in equation 1. That equation gives the likelihood of one raw score conditional on the param-

eters. Let N be the total number of raw scores. x is a $N \times 1$ vector of these raw scores. If all of the x_i's are independent of each other, then the likelihood of the data is

$$p(x|G, \pi, \mu, \sigma^2) = \prod_{i=1}^{N} \sum_{g=1}^{G} \pi_g \phi(x_i|\mu_g, \sigma_g^2), \tag{7}$$

where π, μ, and σ^2 are $G \times 1$ vectors of parameters. I fix G at 13 and estimate the other parameters. The joint distribution of these parameters is

$$f(\pi, \mu, \sigma^2|G, x) \propto p(x|G, \pi, \mu, \sigma^2) h(\pi, \mu, \sigma^2), \tag{8}$$

where $h(\cdot)$ is the conjugate prior.

A.1 Setting the prior

This prior is conjugate, namely,

$$\mu_g \sim \mathcal{N}(\xi_g, \nu_g), \sigma_g^2 \sim IG(\alpha_g, \beta_g), \pi \sim \mathcal{D}(\delta), \tag{9}$$

where $\mathcal{N}(\xi, \nu)$ is the Normal distribution with mean ξ and variance ν. $IG(\alpha, \beta)$ is the Inverse Gamma distribution with mean of $1/[\beta(\alpha - 1)]$. $\mathcal{D}(\delta)$ is the Dirichlet distribution.

I make the prior have very little information.

A.1.1 Component means (μ)

The prior for component means is determined by two vectors: ξ and ν. ξ are the prior means of component means. Since the data is always restricted to the $(0, 100)$ interval, I make the elements of ξ approximately equidistant on that interval. Thus, for $G = 13$, $\xi_g = 7g$.

To reflect that I am very uncertain that these values of ξ correspond to the true component means, I set the prior variance of component means, ν, to the high value of 400.

A.1.2 Component variances (σ^2)

I want the prior standard deviation of all components to be approximately 5. If the expected value of prior standard deviations is 5 and their variance is 4, then the expected value of prior variances is $m = 29$. To make the prior distribution of component variances vague, I follow Carlin & Louis 2000, p. 326 by setting the standard deviation of prior variance equal to its mean m. This means that $\alpha = 3$, $\beta = 1/(2m) = 0.0172$ for all g.

A.1.3 Component probabilities (π)

To make all the components *a priori* equally likely, all the elements of δ must be equal. The elements of δ can be interpreted as the number of fake observations from each component of the mixture. I thus set all elements of δ to 2.

A.2 Estimating posterior distributions

A.2.1 Basic approach

The posterior of equation 8 cannot be calculated analytically. I thus use a technique called Gibbs Sampler to numerically estimate this distribution. Suppose there are R unknown parameters, namely $\Theta = (\theta_1, \theta_2, \dots, \theta_R)$. Let Θ_{-r} be all of the parameters except θ_r. Let x be the data. Suppose that all of the conditional distributions $f(\theta_r|x, \Theta_{-r})$ are known. Let $\theta_r^{(t)}$ be the simulated draw from the posterior of parameter θ_r during the t-th iteration. Then, the Gibbs Sampler algorithm is:

Algorithm 4 (Gibbs Sampler) *1. Set $\Theta^{(0)}$ to some initial values.*

2. *If $\Theta^{(t)}$ is known, sample $\theta_r^{(t+1)}$ from $f(\theta_r|x, \theta_1^{(t+1)}, \dots \theta_{r-1}^{(t+1)}, \theta_{r+1}^{(t)}, \dots \theta_R^{(t)})$.*
3. *Use the draws from the previous step to construct $\Theta^{(t+1)}$ and repeat for $t = 1$ to T.*
4. *Discard $\Theta^{(t)}$ for all $t \leq B$, where $B \ll T$ is the "burn-in" period. The remaining values of $\Theta^{(t)}$ are the simulated draws from the posterior distribution of Θ.*

Let z_i be equal to the component of the mixture to which observation x_i belongs. The unknown parameters in our application are $\Theta = (z, \pi, \mu, \sigma^2)$. Augmenting the unknown parameters with z makes it easy to find the conditional distributions needed by the Gibbs Sampler algorithm. Let $|\dots$ mean conditioning on all other parameters in Θ, the data x, and $G = 13$. The conditional distributions are:

$$\Pr(z_i = g|\dots) \propto \pi_g \phi(x_i|\mu_g, \sigma_g^2), \tag{10}$$

$$\pi|\dots \sim \mathcal{D}(\delta + n), \text{ where } n_g = \#\{i : z_i = g\}. \tag{11}$$

n_g is simply the number of observations in group g according to the parameter z. For writing down the conditional for posterior means μ_g, it is convenient to define two quantities: D_g and d_g.

$$D_g = (n_g \sigma_g^{-2} + \nu_g^{-1})^{-1}, \tag{12}$$

$$d_g = \sigma_g^{-2} \sum_{i:z_i=g} x_i + \nu_g^{-1} \xi_g, \tag{13}$$

$$\mu_g | \ldots \sim \mathcal{N}(D_g d_g, D_g). \tag{14}$$

Finally, the conditional for posterior variances σ_g^2 is

$$\sigma_g^2 | \ldots \sim IG \left(\alpha_g + n_g/2, \left(\beta_g^{-1} + 0.5 \times \sum_{i:z_i=g} (x_i - \mu_g)^2 \right)^{-1} \right). \tag{15}$$

Note that in the expressions for d_g and σ_g^2, the summation is taken over all the observations that belong to group g.

A.2.2 Label switching

Though the Gibbs Sampler algorithm described above works in general, in case of mixture models there is an issue called *label switching*. The problem is that, just following the basic algorithm, it is impossible to identify which component of the mixture a draw is being made from. As a result, posterior densities for all components look the same.

A basic solution to this issue is to impose *identifiability constraints* when they can be found. Fortunately, in the application of grading, there is a very natural constraint, namely,

$$\mu_1 < \mu_2 < \ldots < \mu_G. \tag{16}$$

That is, the mean raw score of F's is less than the mean raw score of $D-$'s, and so on. The Gibbs Sampler draws are post-processed to enforce this constraint.

A.2.3 Initial values

Here, I explain how the initial parameter values $\Theta^{(0)}$ are set. Following Raftery 1996, p. 176, I first sort the data and subdivide it into $G = 13$ groups of equal size. The lowest observations are in group one, the lowest observations that are not in group one are in group two, and so on. I then set the initial means and variances to the sample quantities from the corresponding groups.

As for the initial values of component probabilities, I set them all to $\frac{1}{G}$.

A.2.4 Convergence

Lastly, I have to set the total number of draws T and the burn-in period B. I find that, in this application, convergence is achieved very quickly and so this is not an issue. I set $T = 2000$ and $B = 500$. As an example, see Figure 3. It shows the posterior log likelihood for "Large Class". The graph shows a stable log likelihood, indicating convergence.

A.3 Estimation and performance

I use the average over the Gibbs Sampler draws as point estimates of parameters. This corresponds to a quadratic loss over these parameters.

The algorithms takes about a minute or two on a reasonably modern computer. Compare this to the *twenty minutes* that it took a computer in 1983 to calculate letter grades using the Standard Deviation method (see Spencer 1983). And that was considered an improvement over using a calculator!

	Straight Scale		**Standard Deviation**		**Curve**
	From	To	From	To	% students
F	0.0	60.0	0	$\mu - 1.5 \times \sigma$	6.68%
$D-$	60.0	63.3	$\mu - 1.5 \times \sigma$	$\mu - 1.17 \times \sigma$	5.49%
D	63.3	66.7	$\mu - 1.17 \times \sigma$	$\mu - 0.83 \times \sigma$	8.07%
$D+$	66.7	70.0	$\mu - 0.83 \times \sigma$	$\mu - 0.5 \times \sigma$	10.62%
$C-$	70.0	73.3	$\mu - 0.5 \times \sigma$	$\mu - 0.17 \times \sigma$	12.53%
C	73.3	76.7	$\mu - 0.17 \times \sigma$	$\mu + 0.17 \times \sigma$	13.24%
$C+$	76.7	80.0	$\mu + 0.17 \times \sigma$	$\mu + 0.5 \times \sigma$	12.53%
$B-$	80.0	83.3	$\mu + 0.5 \times \sigma$	$\mu + 0.83 \times \sigma$	10.62%
B	83.3	86.7	$\mu + 0.83 \times \sigma$	$\mu + 1.17 \times \sigma$	8.07%
$B+$	86.7	90.0	$\mu + 1.17 \times \sigma$	$\mu + 1.5 \times \sigma$	5.49%
$A-$	90.0	93.3	$\mu + 1.5 \times \sigma$	$\mu + 1.83 \times \sigma$	3.34%
A	93.3	96.7	$\mu + 1.83 \times \sigma$	$\mu + 2.17 \times \sigma$	1.83%
$A+$	96.7	100.0	$\mu + 2.17 \times \sigma$	100	1.51%

Table 1: Common Straight Scale, Standard Deviation, and Curve scales. For Standard Deviation and Curve, the mean is "set" at C. Straight Scale and Standard Deviation show the scores between which a certain letter grade is assigned. Curve shows the percent of students that receive a certain letter grade. μ is the estimated mean of raw scores; σ is the estimated standard deviation.

B Tables and Figures

Component	LG	Ordinate	Probability	Cumulative Pr
1	F	0	0.0%	0.0%
2	$D-$	1.84×10^{-12}	0.0%	0.0%
3	D	1.22×10^{-7}	0.0%	0.0%
4	$D+$	1.06×10^{-4}	0.7%	0.7%
5	$C-$	3.08×10^{-3}	20.2%	20.9%
6	C	5.76×10^{-3}	37.8%	58.6%
7	$C+$	4.38×10^{-3}	28.7%	87.3%
8	$B-$	1.63×10^{-3}	10.7%	98.0%
9	B	2.69×10^{-4}	1.8%	99.8%
10	$B+$	2.99×10^{-5}	0.2%	100.0%
11	$A-$	4.84×10^{-6}	0.0%	100.0%
12	A	6.30×10^{-7}	0.0%	100.0%
13	$A+$	6.30×10^{-10}	0.0%	100.0%

Table 2: See figure 1 and examples 1 and 3. LG is the letter grade corresponding to a Component. Ordinate is the ordinate at $x = 60$. Probability is the probability that $x = 60$ belongs to the given component. Cumulative Pr is the probability that $x = 60$ belongs to the given or lower component.

	MCG				Grader	
	neutral		lenient			
	to	percent	to	percent	to	percent
F	11.0	1.9%	11.0	1.9%	39.0	5.6%
$D-$	24.3	1.6%	24.3	1.6%		
D	37.0	1.9%	35.0	1.6%	49.3	4.0%
$D+$	47.0	3.5%	46.0	3.5%		
$C-$	*54.8*	4.8%	52.9	3.5%	54.0	3.7%
C	61.4	9.6%	*59.8*	8.8%	59.3	7.0%
$C+$	66.3	9.9%	65.0	9.6%	64.3	7.8%
$B-$	*69.8*	5.1%	68.3	6.7%	69.3	9.9%
B	**73.5**	8.6%	**72.3**	6.7%	**74.3**	12.0%
$B+$	**78.5**	18.5%	**77.0**	15.5%	79.3	19.3%
$A-$	83.5	18.7%	82.3	20.1%	86.3	20.9%
A	89.5	10.2%	87.5	12.6%	94.3	9.1%
$A+$	99.0	5.9%	99.0	8.0%	99.0	0.8%

Table 3: See section 6.2 ("Large Class"). Maximum score for each grade and percent of students receiving that grade.

	Neutral CC	Lenient CC	$\mathbf{R}_r^2$
Straight Scale	1.12	1.67	83.7%
SD (mean C)	0.81	1.22	92.1%
SD (mean B)	0.16	*0.16*	*99.3%*
Adjusted Scale	1.06	1.58	85.0%
Maximum	0.75	1.12	93.0%
MCG	*0.14*	0.19	*99.3%* / 99.1%

Table 4: See section 6.2. Performance of selected grading methods for "Large Class". First two columns show Class Loss (CC). In the first column, Leniency factor is neutral; in the second, it is lenient. Third column shows raw coefficient of dermination (R_r^2). For MCG, the first R_r^2 value is neutral, the second is lenient.

	MCG				**Grader**	
	neutral		**lenient**			
	to	**percent**	**to**	**percent**	**to**	**percent**
F						
D−	39.7	0.8%				
D			39.7	0.8%		
D+	50.6	1.6%	50.6	1.6%		
C−	55.8	10.9%	53.8	3.1%	55.1	9.3%
C	58.8	6.2%	57.1	9.3%	61.4	20.9%
C+	*62.0*	14.7%	60.5	11.6%	63.1	10.1%
B−	**65.2**	18.6%	*63.5*	17.1%	64.2	8.5%
B	*68.3*	17.8%	**66.6**	17.1%	**67.4**	17.8%
B+	71.7	15.5%	70.0	20.9%	69.1	7.8%
A−	74.8	4.7%	73.7	7.8%	70.6	7.8%
A	*78.3*	8.5%	77.1	9.3%	78.3	17.1%
A+	83.7	0.8%	83.7	1.6%	83.7	0.8%

Table 5: See section 6.3 ("Medium Class"). Maximum score for each grade and percent of students receiving that grade.

	Neutral CC	Lenient CC	$\mathbf{R}_r^2$
Straight Scale	1.87	2.81	58.7%
SD (mean C)	0.81	1.22	91.1%
SD (mean B)	0.25	0.27	98.8%
Adjusted Scale	0.60	0.90	94.8%
Maximum	0.60	0.90	94.8%
MCG	*0.18*	*0.24*	*99.1%* / *99.0%*

Table 6: See section 6.3. Performance of selected grading methods for "Medium Class".

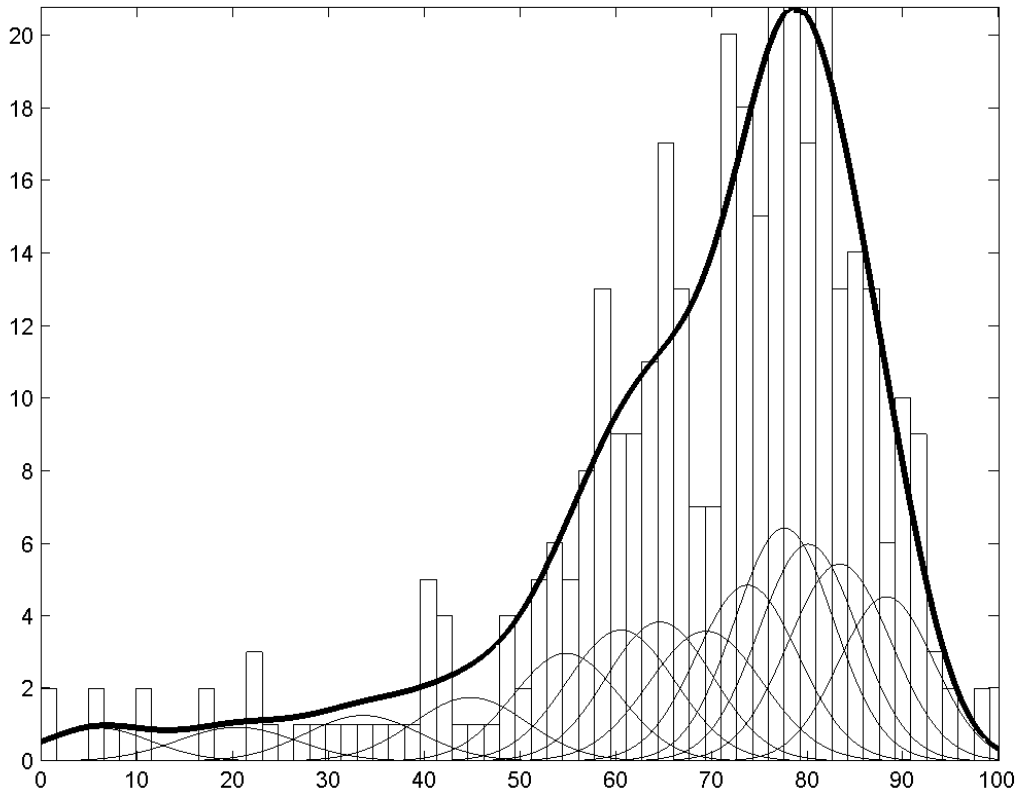

Figure 1: A histogram of raw scores from "Large Class", along with an estimated distribution of those scores. Upper line shows the combined distribution. Lower lines show components.

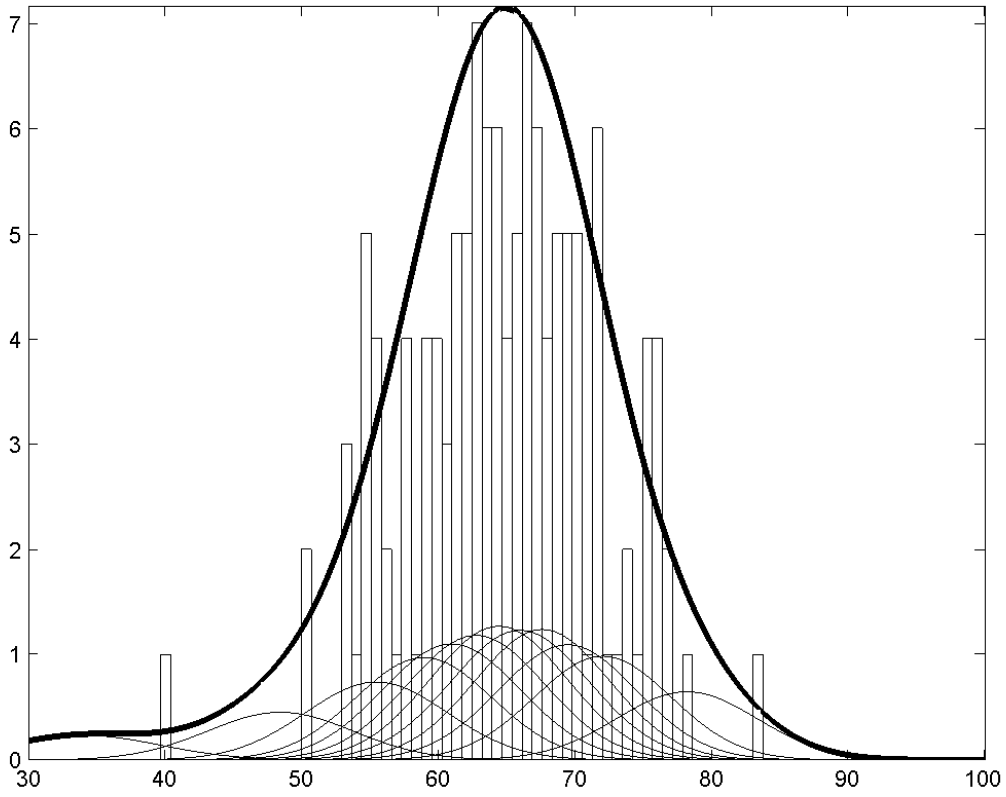

Figure 2: A histogram of raw scores from "Medium Class", along with an estimated distribution of those scores.

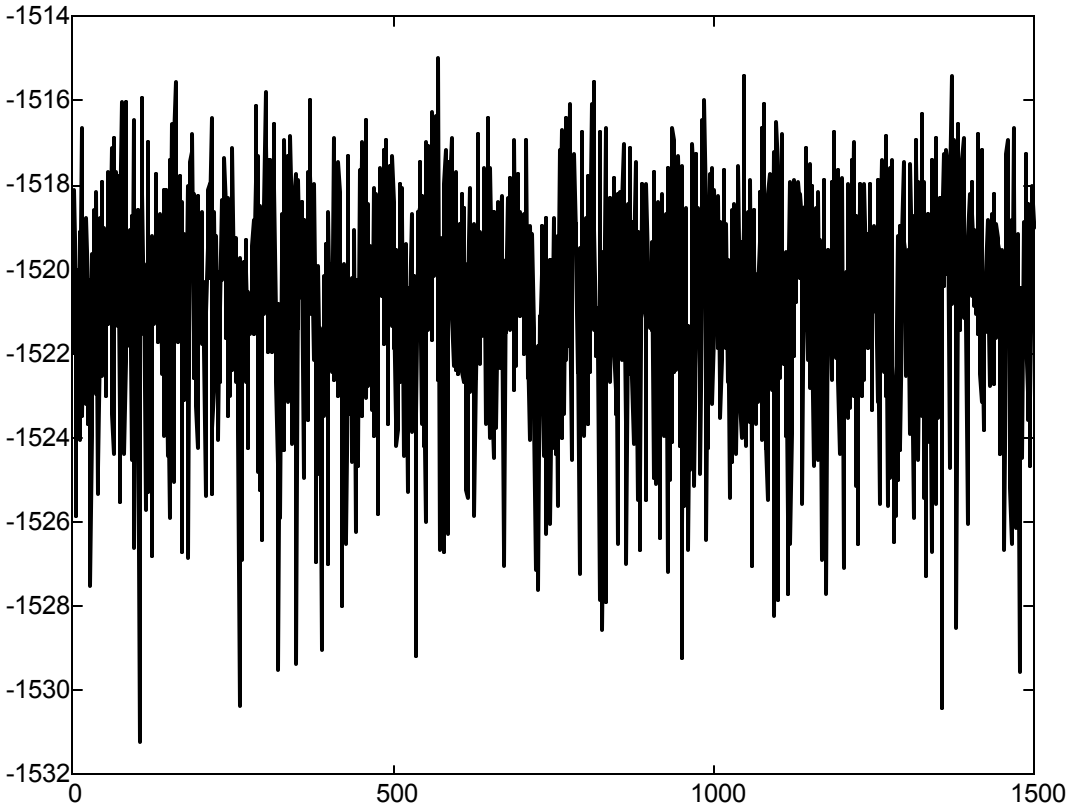

Figure 3: Posterior log likelihood for estimating the model for "Large Class". The stable likelihood indicates convergence.