Bayesian Shrinkage

The Collection — Rating Adjustment

Hierarchical Model, Derivation, and Application to Codex Data

Jure Skarabot • MMXXVI

A. Setting and Notation

The codex contains 118 wines partitioned into K groups (countries, or analogously grape categories). Let group i have n_i wines, each an independent rating drawn from the group's underlying distribution, with sample mean

$$\bar{x}_i \;=\; \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij}.$$

We want an estimator of each group's true mean rating. The naïve choice is the sample mean itself: the maximum likelihood estimator under independent sampling, which is unbiased. But when some n_i are small, the sample-mean estimator has large variance — a country with $n = 1$ has variance equal to the full population variance — and the resulting rankings are dominated by sampling noise rather than by real differences between groups.

Bayesian shrinkage resolves this by imposing a hierarchical model: the group means are themselves drawn from a common distribution, and each group's sample mean is combined with that population-level information to produce a posterior estimate. The result is a point estimator that is biased but has smaller mean squared error than the MLE across every realistic configuration of the true means.

A.1 The Normal–Normal Hierarchical Model

Assume:

$$\begin{aligned} x_{ij} \mid \theta_i &\sim \mathcal{N}(\theta_i,\, \sigma^2), \quad \text{independent across } i,\, j \\ \theta_i &\sim \mathcal{N}(\mu,\, \tau^2), \quad \text{independent across } i. \end{aligned}$$

Here the within-group variance (how much ratings vary within a single country) is denoted by one parameter, the between-group variance (how much true country means vary) by another, and the grand mean by a third. This is the simplest non-trivial hierarchical model and suffices for rating data bounded roughly in [8.9, 10.0], where normality is a reasonable working assumption.

Given the model, the group's sample mean is a sufficient statistic for the true mean, and its distribution conditional on the true mean is

$$\bar{x}_i \mid \theta_i \;\sim\; \mathcal{N}\!\left(\theta_i,\, \sigma^2/n_i\right).$$

B. Posterior Derivation

B.1 Conjugacy and the Posterior Mean

Because the Normal family is self-conjugate, the posterior for the true mean is also Normal. Applying Bayes' rule with a Normal prior and a Normal likelihood, the posterior precision is the sum of prior and likelihood precisions:

$$\frac{1}{\operatorname{Var}(\theta_i \mid \bar{x}_i)} \;=\; \frac{1}{\tau^2} + \frac{n_i}{\sigma^2}.$$

Vinotheca • *published under CC BY-NC 4.0*

The posterior mean is the precision-weighted average of the prior mean and the data mean:

$$\mathbb{E}[\theta_i \mid \bar{x}_i] \;=\; \frac{\mu/\tau^2 + n_i\,\bar{x}_i/\sigma^2}{1/\tau^2 + n_i/\sigma^2}.$$

Multiplying numerator and denominator by the within-group variance and rearranging:

$$\hat{\theta}_i \;=\; \frac{n_i\,\bar{x}_i + (\sigma^2/\tau^2)\,\mu}{n_i + \sigma^2/\tau^2}.$$

Define the shrinkage constant as the ratio of within- to between-group variance:

$$m \;=\; \sigma^2/\tau^2.$$

The estimator takes the clean closed form

$$\boxed{\hat{\theta}_i \;=\; \frac{n_i\,\bar{x}_i + m\,\mu}{n_i + m}}.$$

This is the formula implemented on the site. It is the posterior mean under the Normal–Normal hierarchical model; equivalently, it is the Bayes estimator under squared-error loss. As the sample size grows, the estimator converges to the raw mean; as it vanishes, the estimator converges to the prior mean. The ratio m has an intuitive reading: it is the number of observations whose evidence is worth, in aggregate, the same as the prior — a prior "sample size."

B.2 Equivalent Weighted-Average Form

Define

$$\alpha_i \;=\; \frac{n_i}{n_i + m}.$$

Then the estimator becomes

$$\hat{\theta}_i \;=\; \alpha_i\,\bar{x}_i + (1 - \alpha_i)\,\mu.$$

This makes the shrinkage explicit: the estimate is a convex combination of the group mean and the grand mean, with the weight on the group mean growing with sample size. The factor $1 - \alpha_i$ is the *shrinkage fraction*. When the sample size equals the shrinkage constant m, the estimate sits exactly halfway between the group mean and the grand mean.

C. Empirical Bayes Estimation from Codex Data

C.1 Variance Components by Method of Moments

The formula requires both variance components. In a fully Bayesian treatment these would have their own priors; the *empirical Bayes* approach estimates them from the data by method of moments.

With the current 118 wines:

$$\begin{aligned}
\hat{\mu} &= 9.516,\\
\hat{\sigma}^2 &= \mathrm{MSW} = 0.0616,\\
\hat{\tau}^2 &= \max\!\big(0,\; (\mathrm{MSB} - \mathrm{MSW})/\bar{n}\big) = 0.0039,
\end{aligned}$$

where MSW is the mean squared error within countries, MSB is the mean squared error between country means (weighted by sample size), and the average sample size across countries with at least two wines is 10.45. Singletons are excluded from the between-group variance estimation because their contribution to the between-group sum of squares is indistinguishable from within-group noise.

The empirical-Bayes shrinkage constant is

$$\hat{m}_{\mathrm{EB}} \;=\; \hat{\sigma}^2/\hat{\tau}^2 \;\approx\; 15.69.$$

C.2 Choice of a Fixed Shrinkage Constant

The site uses a fixed $m = 5$ rather than the empirical-Bayes estimate. With the empirical-Bayes estimate substantially larger than 5, it would shrink ranks more aggressively than the site currently does — reflecting that within-country variance greatly exceeds between-country variance, so most of the observed spread in raw country means is sampling noise rather than real differences in true group means. Two practical considerations argue against simply adopting the empirical-Bayes value.

Stability of the estimate. The between-group variance is itself estimated from only 11 groups with at least two wines. Its sampling distribution is wide enough that the empirical-Bayes estimate could easily double or halve as the codex grows by a dozen wines. A fixed m produces a stable ranking that is not sensitive to small changes in the data.

Visual legibility. At an empirical-Bayes value near 15.7, small-n countries are shrunk so close to the grand mean that the chart becomes visually flat and conveys little. $m = 5$ preserves more of the raw signal while still correcting for the most obvious small-sample distortions. This is a deliberate bias-for-legibility trade-off.

D. Worked Example

Applying the estimator with $m = 5$ and grand mean 9.516 to the current country data:

Country (i)	n_i	$\bar{x}_i$	α_i	$\hat{\theta}_i$	shift
France	32	9.591	0.865	9.581	−0.010
Italy	26	9.485	0.839	9.490	+0.005
USA	25	9.548	0.833	9.543	−0.005
Germany	8	9.537	0.615	9.529	−0.008
Austria	5	9.480	0.500	9.498	+0.018
Spain	5	9.280	0.500	9.398	+0.118
South Africa	3	9.333	0.375	9.448	+0.114
Argentina	3	9.767	0.375	9.610	−0.157
New Zealand	3	9.333	0.375	9.448	+0.114
Australia	3	9.433	0.375	9.485	+0.052
Chile	2	9.750	0.286	9.583	−0.167
Slovenia	1	8.900	0.167	9.413	+0.513
Greece	1	9.200	0.167	9.463	+0.263
Croatia	1	9.400	0.167	9.497	+0.097

The α column is the effective weight on each country's own data. France ($n = 32$) has α close to 0.86, so its estimate barely moves. Countries with $n = 1$ have α close to 0.17, meaning about 83% of their estimate comes from the grand mean — a direct reflection of how little we know about them. Notice that the shrinkage magnitude depends on both the sample size *and* how far the raw mean sits from the grand mean: Argentina's three wines average far above the grand mean, so it gets pulled visibly toward it; Austria sits near the grand mean and barely moves despite a similar sample size.

E. Properties of the Estimator

Under the assumed model, the estimator has three notable properties.

E.1 Bias Toward the Grand Mean

The bias vanishes as the sample size grows. Explicitly, the expected value of the estimator minus the true mean equals the product of the shrinkage fraction and the distance between the true

mean and the grand mean:

$$\mathbb{E}[\hat{\theta}_i] - \theta_i \;=\; (1 - \alpha_i)(\mu - \theta_i).$$

E.2 Lower Variance than the MLE

The variance of the estimator conditional on the true mean is the MLE's variance multiplied by the square of α:

$$\mathrm{Var}(\hat{\theta}_i \mid \theta_i) \;=\; \alpha_i^2 \cdot \frac{\sigma^2}{n_i}.$$

Since α_i is strictly less than one, this is a meaningful reduction.

E.3 Admissibility

There exists no estimator that dominates the posterior mean in mean squared error uniformly over all possible configurations of the true group means. This is a standard decision-theoretic result in the Normal means problem.

E.4 Implications for Ranking

For group rankings — the use case here — what matters is that the posterior mean is the Bayes estimator under squared-error loss, and that it orders groups in a way that accounts for sampling uncertainty. A country ranked high by the estimator is high either because its true mean is high, or because its sample size is large enough that the sample mean is trustworthy; either is a defensible reason to display it high.

F. Limitations

F.1 Bounded and Censored Ratings

The normality assumption is a working convenience. Ratings are bounded and right-censored near 10.0 — there is no rating above 10.0, and my own 9.9s are de facto truncated draws from whatever unbounded latent variable they represent. For bounded ratings with potential ceiling effects, a Beta-Binomial or ordered-logit model would be more principled but would not change the qualitative ranking meaningfully at current sample sizes.

F.2 Exchangeability Across Groups

The model assumes the true group means are exchangeable across countries, which is a convenient fiction. A Burgundy and a Mosel wine are drawn from visibly different underlying distributions. A more refined model would introduce further hierarchy — continent, climate zone, price tier — at the cost of requiring substantially more data to identify the additional variance components.

F.3 Selection in the Underlying Sample

Finally, the ratings themselves are not independent samples from a well-defined population. They reflect my purchase patterns, which are non-random — I buy what I expect to enjoy — and my tasting context, which varies. No amount of statistical machinery fixes that; shrinkage only addresses the sample-size problem conditional on whatever selection process generated the data in the first place.

G. References

- Efron, B. & Morris, C. (1975). "Data Analysis Using Stein's Estimator and Its Generalizations." *Journal of the American Statistical Association* 70 (350): 311–319.

- Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*, 3rd ed., chapter 5.
- Robinson, D. (2017). *Introduction to Empirical Bayes: Examples from Baseball Statistics.*