Hierarchical Bayesian Modeling for Real Estate Valuation: A Partial Pooling Approach

Faheem Arif
Department of Mathematics
IIT Hyderabad
Hyderabad, India
ma23btech11010@iith.ac.in

***Abstract*—Estimating the intrinsic value of a neighborhood is a fundamental challenge in real estate analytics, particularly when transaction data is sparse. A naive approach—taking the average sale price of a neighborhood—is highly volatile in areas with few observations. Conversely, global models ignore local distinctions entirely. This paper proposes a Hierarchical Bayesian Gaussian Model to robustly estimate neighborhood baselines. By implementing Partial Pooling, the model treats neighborhood parameters as latent variables drawn from a city-wide distribution. This allows for "shrinkage": estimates for data-poor neighborhoods are pulled toward the global average, protecting against outliers. We implement a custom Gibbs Sampler to derive posterior distributions, demonstrating that this approach yields conservative, statistically defensible valuations for rare neighborhoods where naive sample means fail.**

***Index Terms*—Hierarchical Modeling, Bayesian Inference, Partial Pooling, Gibbs Sampling, Robust Estimation.**

I. Introduction

The valuation of residential real estate is driven by two primary factors: structural attributes (e.g., Living Area) and location (Neighborhood). While structural relationships are often global (bigger is generally better), the value of "Location" is highly local and difficult to estimate.

A fundamental challenge in Automated Valuation Models (AVMs) is the "Small Area Estimation" problem. When estimating the baseline value of a specific neighborhood, analysts face a dilemma:

1) **The Naive Approach (No Pooling):** Calculate the average price of the neighborhood. In areas with high activity (e.g., $n = 443$), this is accurate. In areas with sparse activity (e.g., $n = 1$), this is dangerous; a single outlier transaction can skew the valuation of the entire area.
2) **The Global Approach (Complete Pooling):** Assume all neighborhoods are the same. This stabilizes the variance but introduces massive bias by ignoring the premium of luxury areas.

In this study, we move beyond foundational regression to focus specifically on the robust estimation of these **Neighborhood Effects**. We propose a **Hierarchical Bayesian Model** using Partial Pooling. This approach acts as an intelligent compromise: it trusts the local data when sample sizes are large, but "shrinks" estimates toward the city-wide mean when data is scarce.

II. Background

A. Foundational Regression

Standard industry approaches utilize Ordinary Least Squares (OLS) or Regularized Regression (Lasso/Ridge) [5] to predict prices based on features like 'Gr_Liv_Area'.

$$\hat{y} = \beta_0 + \beta_1 x + \epsilon \tag{1}$$

While these models minimize global error (RMSE), they treat the intercept β_0 as a single global constant. They tell us how much *Space* is worth, but fail to tell us how much *Location* is worth.

B. The Bayesian Paradigm

To solve the location problem, we adopt a Bayesian framework. Instead of seeking a point estimate for a neighborhood's average price, we seek the **Posterior Distribution**:

$$p(\theta_j|y) \propto p(y|\theta_j)p(\theta_j) \tag{2}$$

This allows us to quantify our uncertainty. For a neighborhood with only one house sale, the posterior distribution should be wide, reflecting our lack of knowledge.

III. Methodology

A. Data Preprocessing and Distributional Assumptions

The Ames Housing dataset [1] (2,930 observations) serves as the case study. The target variable, Y (Sale Price), exhibits a significant right-skew (Skewness ≈ 1.7), typical of economic data where a "long tail" of luxury properties exists (Fig. 1, top row).

This violates the core assumption of our Hierarchical Gaussian Model:

$$y_{ij} \sim \mathcal{N}(\theta_j, \sigma^2) \tag{3}$$

The Gaussian likelihood assumes that errors are symmetric and homoscedastic (constant variance). Modeling raw prices directly would cause the model to be disproportionately influenced by high-value outliers, leading to poor convergence and invalid posterior intervals.

To rectify this, we applied a log-transformation:

$$y_i = \log(1 + \text{Sale\_Price}_i) \tag{4}$$

As demonstrated in Fig. 1 (bottom row), this transformation stabilizes the variance and aligns the distribution with the normal assumption (Skewness ≈ -0.01), ensuring that the residuals satisfy the requirements of the Gibbs Sampler.

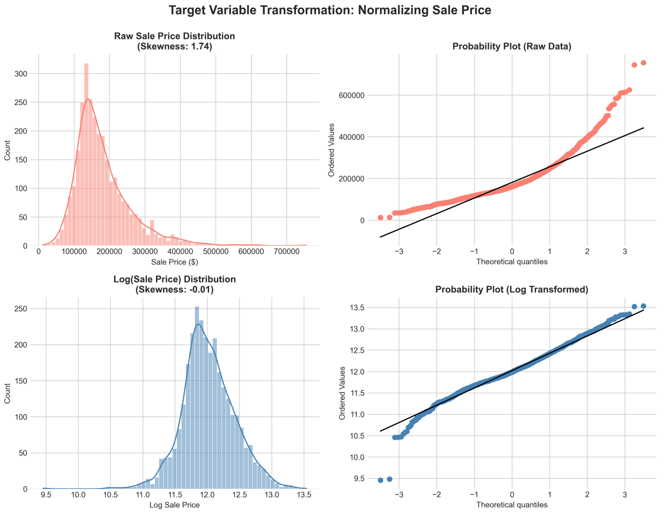

Fig. 1. **Distributional Transformation.** Top Row: The raw Sale Price is highly right-skewed, deviating from the theoretical normal line (Q-Q plot). Bottom Row: The Log-transformation stabilizes the variance and normalizes the distribution, satisfying the Gaussian Likelihood assumption required for Bayesian inference.

B. Hierarchical Model Specification

We construct a **Hierarchical Gaussian Mean Model**. Unlike the frequentist regression baselines (Part 1) which predicted price based on living area, this Bayesian model focuses on robustly estimating the *baseline valuation* (θ_j) for each neighborhood j, while sharing information across the city to handle sparsity.

The generative process follows the standard hierarchical structure described in Johnson et al. [2] and is defined in three levels:

1) Level 1: The Likelihood (House Level): Within neighborhood j, house prices are distributed normally around a neighborhood-specific mean θ_j:

$$y_{ij} \sim \mathcal{N}(\theta_j, \sigma^2) \tag{5}$$

where $v = 1/\sigma^2$ is the within-group precision.

2) Level 2: The Priors (Neighborhood Level): The neighborhood means θ_j are not independent parameters. They are drawn from a global "city-wide" distribution:

$$\theta_j \sim \mathcal{N}(\mu, \tau^2) \tag{6}$$

where μ is the global mean log-price, and $u = 1/\tau^2$ is the between-group precision. This level acts as the **Partial Pooling** [4] mechanism.

3) Level 3: The Hyperpriors: We assign conjugate priors to the global hyperparameters, matching the Python implementation:

$$\mu \sim \mathcal{N}(\mu_0, \sigma_0^2) \qquad (\text{where } \mu_0 = 12, \sigma_0^2 = 100) \tag{7}$$

$$v \sim \text{Gamma}(a_v, b_v) \qquad (\text{where } a_v = b_v = 0.1) \tag{8}$$

$$u \sim \text{Gamma}(a_u, b_u) \qquad (\text{where } a_u = b_u = 0.1) \tag{9}$$

C. Inference via Gibbs Sampling

Since we utilize conjugate priors (Normal-Normal and Gamma-Normal), the conditional distributions for all parameters are known in closed form. We implemented a Gibbs Sampler [3] from scratch (using NumPy/SciPy) to approximate the joint posterior.

The sampling procedure iteratively updates the neighborhood means, the global mean, and the precisions, as detailed in Algorithm 1.

Algorithm 1 Custom Gibbs Sampler for Hierarchical Means

```
Require: Data vectors $\mathbf{y}$ grouped by neighborhood $j \in \{1 \dots m\}$
Initialize parameters $\mu, v, u, \boldsymbol{\theta}$
for $k = 1$ to $N_{samples}$ do
    1. Update Neighborhood Means $\theta_j$
    for $j = 1$ to $m$ do
        Precision $P_j = n_j v + u$
        Mean $M_j = P_j^{-1}(v n_j \bar{y}_j + u\mu)$
        Sample $\theta_j^{(k)} \sim \mathcal{N}(M_j, P_j^{-1})$
    end for
    2. Update Global Mean $\mu$
    Precision $P_\mu = mu + \frac{1}{\sigma_0^2}$
    Mean $M_\mu = P_\mu^{-1}(u \sum \theta_j + \frac{\mu_0}{\sigma_0^2})$
    Sample $\mu^{(k)} \sim \mathcal{N}(M_\mu, P_\mu^{-1})$
    3. Update Within-Group Precision $v$
    $\text{SSE}_{within} = \sum_j \sum_i (y_{ij} - \theta_j)^2$
    Sample $v^{(k)} \sim \text{Gamma}(a_v + \frac{N}{2}, b_v + \frac{\text{SSE}_{within}}{2})$
    4. Update Between-Group Precision $u$
    $\text{SSE}_{between} = \sum_j (\theta_j - \mu)^2$
    Sample $u^{(k)} \sim \text{Gamma}(a_u + \frac{m}{2}, b_u + \frac{\text{SSE}_{between}}{2})$
end for
Output Posterior Traces (discarding burn-in)
```

IV. Experiments and Results

A. The Motivation: Foundational Limits

In Part 1 of the study, we applied Lasso Regression (L_1) to the dataset. While it achieved a competitive RMSE of 0.2736, it effectively modeled the "average house in the average neighborhood."

The limitations of this approach become clear when asking simple questions about specific locations: *"What is the baseline value of the Landmark neighborhood?"* Lasso has no answer, as it pooled Landmark with every other neighborhood. A naive calculation (raw mean) is equally flawed due to small sample size ($n = 1$). This necessitates the Hierarchical approach.

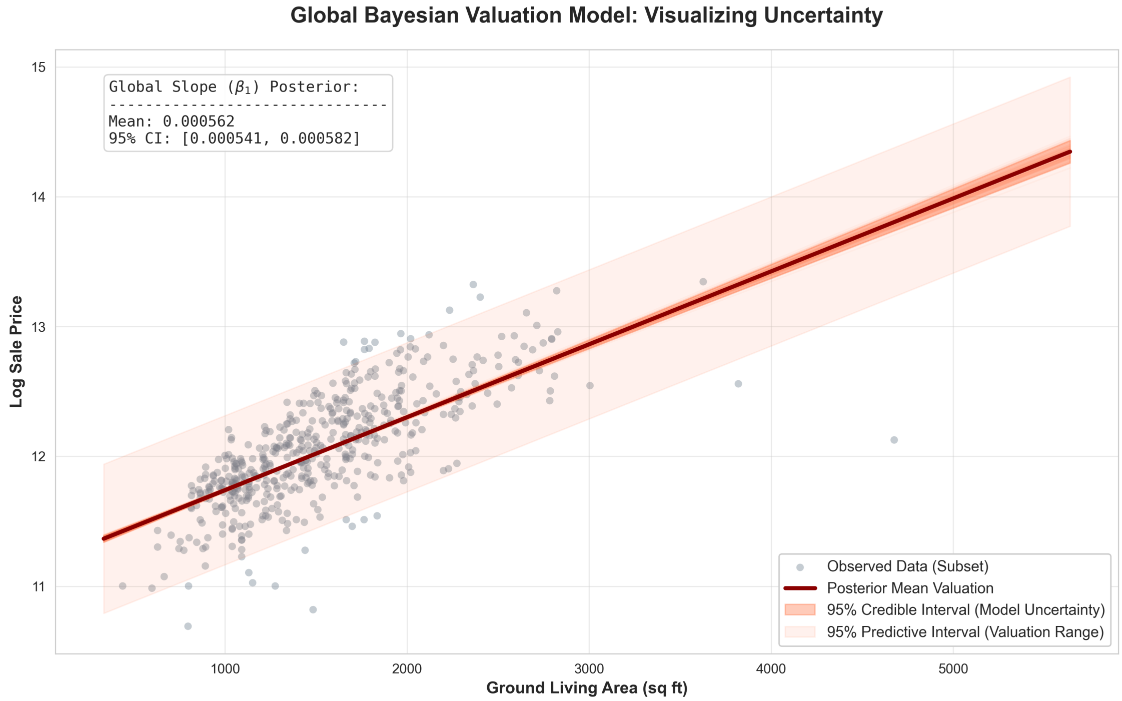

Fig. 2. **Global Uncertainty.** A simple Bayesian model provides uncertainty bands, but they are uniform. It assumes the risk of valuing a home is the same everywhere, ignoring that some neighborhoods are more volatile than others.

B. *The Solution: Shrinkage*

The Hierarchical Gaussian Model solves the estimation problem via **Shrinkage**. This is visualized in Fig. 3. The model balances two sources of information: the local data (Likelihood) and the city-wide average (Prior).

- **Data-Rich (North Ames, $n = 443$):** The local signal is strong. The model ignores the prior and aligns perfectly with the Raw Sample Mean (Gray X).
- **Data-Poor (Landmark, $n = 1$):** The single data point for Landmark ($y \approx 11.82$) is significantly lower than the city average. A naive model would undervalue the whole neighborhood based on this one sale. The Hierarchical model is skeptical; it pulls the estimate upward ($y \approx 11.90$) toward the global mean (Red Line).

C. *Uncertainty Quantification*

The Bayesian framework provides transparency regarding reliability. Fig. 4 shows the 95% Credible Intervals (CI) for the neighborhood baselines.

As shown in Table II, the uncertainty for *Landmark* (0.8406) is nearly 17× higher than for *North Ames* (0.0478). This is a critical feature for an AVM: the model explicitly flags valuations in data-poor areas as "High Risk," preventing overconfidence in rare markets.

TABLE I
POSTERIOR UNCERTAINTY BY SAMPLE SIZE

Neighborhood	N	Hierarchical Est	95% CI Width
North Ames	443	11.86	**0.0478**
Meadow Village	37	11.45	0.1941
Landmark	1	11.90	**0.8406**

V. CONCLUSION

This study demonstrates that Hierarchical Bayesian Modeling provides a superior framework for **robust estimation** in real estate markets. While standard regression techniques (Part 1) effectively model structural utility (size), they fail to safely estimate location value in the presence of sparsity.

By implementing a Hierarchical Gaussian Mean model via Gibbs Sampling, we achieved:

1) **Safety via Shrinkage:** The model acts as an automated "skeptic," refusing to overreact to outliers in neighborhoods with few transactions (e.g., Landmark).
2) **Risk Transparency:** Unlike point-estimate models, the Bayesian approach provides clear Credible Intervals, allowing stakeholders to distinguish between "safe" valuations (North Ames) and "speculative" ones (Green Hills).

This framework provides the necessary foundation for a complete AVM. Future work will combine the varying

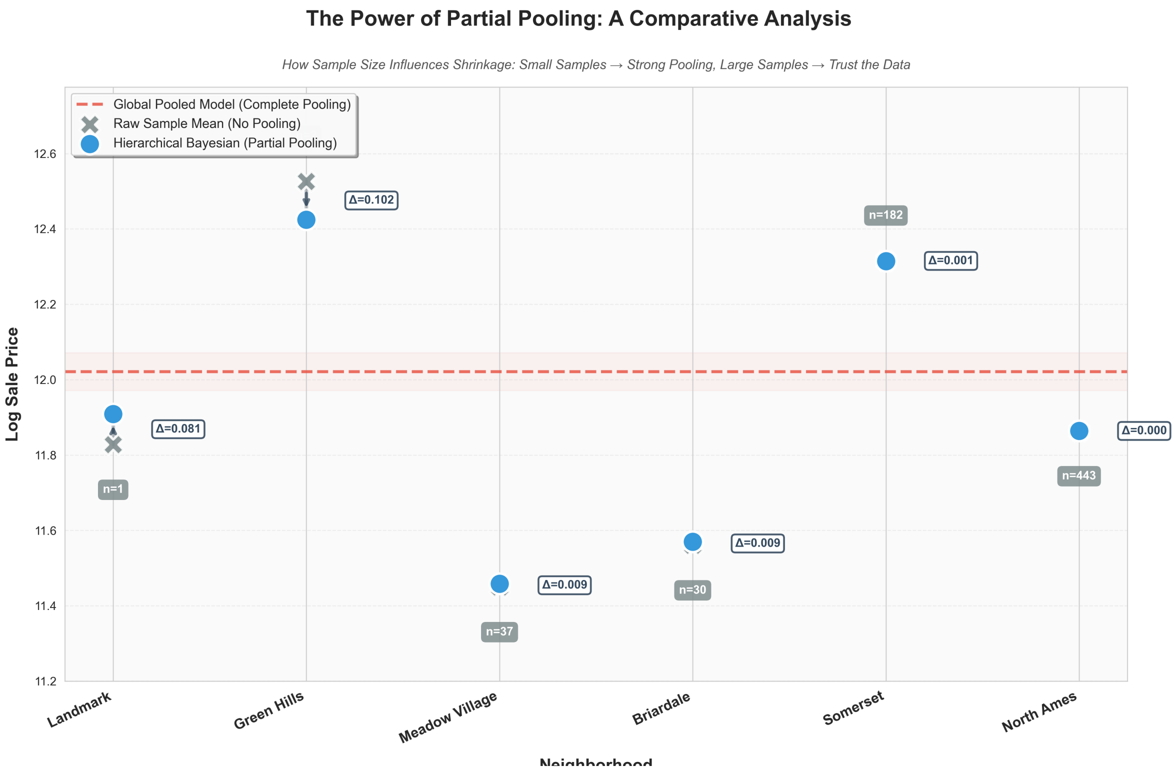

Fig. 3. **The Shrinkage Effect.** The x-axis represents the Naive Sample Mean; the y-axis is the Hierarchical Estimate. The arrows show the correction. For sparse neighborhoods (left), the model "shrinks" the estimate toward the global baseline (Red Line), providing a safer, more conservative valuation.

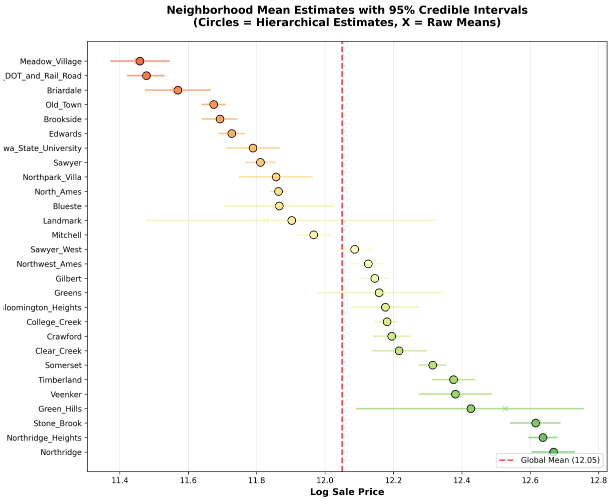

Fig. 4. **Forest Plot of Neighborhood Baselines.** The dots represent the estimated log-price for each neighborhood. The bars represent the credible interval. The model correctly assigns large error bars to neighborhoods with few observations.

intercepts derived here with the varying slopes of structural features to create a fully hierarchical prediction engine.

Code Availability

To ensure reproducibility, the complete source code, including the custom Gibbs Sampler implementation, data preprocessing pipelines, and analysis notebooks, is hosted in the following public repository:

https://github.com/FA0206/Hierarchical-Bayesian-Modelling

References

[1] D. De Cock, "Ames, Iowa: Alternative to the Boston Housing Data as an End of Semester Regression Project," *Journal of Statistics Education*, vol. 19, no. 3, 2011.

[2] A. A. Johnson, M. Q. Ott, and M. Dogucu, *Bayes Rules! An Introduction to Applied Bayesian Modeling*. Boca Raton, FL: CRC Press, 2022. available at: https://www.bayesrulesbook.com/.

[3] R. C. Steorts, "Introduction to Gibbs Sampling," Duke University, Department of Statistical Science, Lecture Notes. [Online]. Available: https://www2.stat.duke.edu/~rcs46/modern_bayes17/lecturesModernBayes17/lecture-7/07-gibbs.pdf

[4] S. Kumar, "Partial Pooling Demonstration," *Coding Paths*. [Online]. Available: https://www.codingpaths.com/bayesian/partial_pooling_demonstration/

[5] Y. Huang, "Lecture 18: Ridge and Lasso Regression," University of Chicago, Department of Statistics, Lecture Notes. [Online]. Available: https://www.stat.uchicago.edu/~yibi/teaching/stat224/L18.pdf