Proceedings of the 4th International Conference on Applied Physics and Mathematical Modeling
DOI: 10.54254/2753-8818/2026.37886

Inferring Energy Gaps in a Three-Level System Using Multinomial Logistic Regression

Huining Su

Beijing No. 35 High School International Department, Beijing
15910634175@139.com

Abstract. Multinomial logistic regression (MNL) and the canonical distribution share an exponential probability structure, allowing regression coefficients to represent energy differences. This study examines the reliability of that correspondence for an ideal, nondegenerate three-level system with dimensionless energies 0, 1 and 2. Independent categorical observations were simulated across four temperature ranges, three sample sizes and three symmetric state-misclassification rates. Two estimators were compared over 200 repetitions per condition: unrestricted MNL and MNL with physically specified zero intercepts. Performance was assessed using energy-gap bias, root mean square error, interval coverage and predictive divergence at unobserved temperatures. With 500 observations per temperature, the constrained upper-gap RMSE was 0.061 in the intermediate range, compared with 0.222 for unrestricted MNL. At 2,000 observations per temperature, 10% misclassification reduced the constrained mean upper-gap estimate from 2.001 to 1.582, with zero empirical interval coverage. The findings distinguish accurate probability prediction from reliable physical inference and demonstrate why measurement calibration must accompany increased sampling. All numerical results are computational; the study evaluates estimation under stated assumptions rather than experimentally validating canonical statistics.

Keywords: multinomial logistic regression, Boltzmann distribution, energy-gap estimation, statistical physics, measurement error

1. Introduction

Discrete-choice methods estimate the probabilities of mutually exclusive outcomes. McFadden established a foundational connection between qualitative choice and logit modelling [1]. Although often applied to behavioural decisions, the exponential normalisation underlying MNL also appears in statistical mechanics. Jaynes connected equilibrium statistical mechanics with information-theoretic inference through the maximum-entropy principle [2]. This shared structure motivates a physical inverse problem: can observed state populations reveal an unknown energy spectrum?

Thermometry provides a related context. Correa et al. showed that the energy-level structure of a quantum probe affects its temperature sensitivity [3]. Hovhannisyan et al. examined how restricted measurement resolution influences thermometric precision [4]. Those studies primarily infer temperature from systems with specified physical properties. The present study reverses the

© 2026 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2753-8818/2026.37886

unknowns: temperature is assumed calibrated, while two energy gaps are estimated from state populations.

The objective is to determine how temperature selection, finite sampling and state misclassification affect MNL-based energy inference. A three-level model provides the smallest setting with more than two categorical outcomes. Comparing free and physically constrained intercepts additionally tests whether correct structural knowledge improves estimation. The contribution is a transparent computational assessment, not a new derivation of the canonical distribution or a claim of experimental spectroscopy.

2. Theoretical framework

2.1 Canonical probabilities and logit coefficients

For a system weakly coupled to a thermal reservoir at equilibrium, the probability of energy level j is given by the canonical distribution [2]:

$$\mathrm{p_j(T)} = \frac{\mathrm{g_j exp[-E_j/(k_B T)]}}{\sum_{\mathrm{k=0}}^{2} \mathrm{g_k exp[-E_k/(k_B T)]}} \tag{1}$$

Here, E_j is the energy, g_j its degeneracy, k_B the Boltzmann constant and T the absolute temperature. The denominator is the partition function. Taking the logarithm of the probability ratio relative to level 0 eliminates this common normalisation:

$$\ln\left(\frac{\mathrm{p_j}}{\mathrm{p_0}}\right) = \ln\left(\frac{\mathrm{g_j}}{\mathrm{g_0}}\right) - \frac{\mathrm{E_j - E_0}}{\mathrm{k_B}} \cdot \frac{1}{\mathrm{T}} \tag{2}$$

An MNL model has log probability ratios $a_j + b_j x$. Setting $x = 1/T$ therefore identifies the intercept with $\ln(g_j/g_0)$ and the slope with $-(E_j - E_0)/k_B$. For known nondegenerate levels, both intercepts equal zero. Adding the same constant to every energy leaves the probabilities unchanged, so only energy differences are identifiable.

For numerical work, let E_{ref} be a fixed, known energy unit, $\theta = k_B T / E_{\mathrm{ref}}$ and $\varepsilon_j = (E_j - E_0)/E_{\mathrm{ref}}$. Then:

$$\mathrm{p_j(\theta)} = \frac{\exp(-\varepsilon_j/\theta)}{1 + \exp(-\varepsilon_1/\theta) + \exp(-\varepsilon_2/\theta)}, \varepsilon_0 = 0 \tag{3}$$

The fitted dimensionless gap is $-b_j$ when $x = 1/\theta$. E_{ref} is a chosen unit, not an estimated gap; dimensional results require a calibrated temperature scale. The generating values are $\varepsilon_1 = 1$ and $\varepsilon_2 = 2$, but the estimators fit both independently without imposing equal spacing. No principal-component transformation is required because inverse temperature already has a direct physical interpretation.

2.2 Information and measurement error

Differentiating the log probability of one observation with respect to the two dimensionless gaps gives the Fisher information matrix:

$$\mathrm{I(\varepsilon;\theta) = \theta^{-2}\,[diag(p*) - p*p*^{T}]} \tag{4}$$

DOI: 10.54254/2753-8818/2026.37886

Here, $p^* = (p_1, p_2)^\top$. This expression describes the zero-intercept model and follows directly from its categorical likelihood. At very low temperatures, excited-state probabilities vanish exponentially. At very high temperatures, the inverse-temperature factor tends to zero. Both limits reduce information about fixed, positive energy gaps, although they need not affect each gap equally. Free intercepts introduce additional uncertainty because temperature-dependent slopes must be distinguished from constant offsets.

State misclassification is represented by probability η of replacing a true label with either alternative, each with probability $\eta/2$. The recorded probabilities are consequently:

$$q_j = (1 - \eta)p_j + (\eta/2)(1 - p_j) = (1 - 3\eta/2)p_j + \eta/2 \tag{5}$$

For positive η, rare states acquire an artificial population floor. These transformed probabilities generally do not retain linear log ratios in inverse temperature. The numerical experiment therefore tests genuine model misspecification as well as finite-sample variation.

3. Methods

3.1 Simulation design

The study specified its aims, generating mechanisms, estimands, methods and performance measures following the ADEMP approach of Morris et al. [5]. Each simulated dataset contained observations at eight geometrically spaced temperatures. Four dimensionless ranges were examined: low, 0.25–0.75; intermediate, 0.75–2; high, 4–10; and broad, 0.25–10. Geometric spacing gave equal spacing on the logarithmic temperature scale.

At every temperature, $n = 100$, 500 or $2,000$ observations were generated, giving 800, 4,000 or 16,000 observations per dataset. Misclassification rates were 0, 0.05 and 0.10. Counts were sampled directly from the multinomial distribution with probabilities in Equation 5, which is distributionally equivalent to independently sampling and misclassifying individual labels. The 36 conditions were each repeated 200 times, producing 7,200 independent datasets and 14,400 model fits. Both estimators used the same dataset within each repetition. NumPy's default random generator used seed 20260908; conditions were traversed in the range order above, then ascending n, η and repetition.

3.2 Estimation and numerical checks

The unrestricted model estimated two intercepts and two slopes; the constrained model fixed the intercepts at zero and estimated two slopes. Both maximised the grouped multinomial likelihood without regularisation, using SciPy's BFGS implementation [6]. Optimisation started at zero, used an analytic gradient, allowed 500 iterations and targeted a gradient tolerance of 10^{-9}. Calculations used NumPy 2.3.5 and SciPy 1.17.0.

Fits were accepted when the maximum absolute gradient of the mean negative log likelihood was below 10^{-6}, the smallest eigenvalue of the total observed-information matrix exceeded 10^{-8}, and all absolute coefficients were below 100. These numerical screens flag unstable fits; they do not constitute a formal separation test. Separation can prevent finite maximum-likelihood estimates in multinomial regression [7]. All 14,400 fits passed the specified screens; none were excluded.

Standard errors were calculated from the inverse analytic observed-information matrix. Nominal 95% Wald intervals used the estimate plus or minus 1.96 standard errors. Under misclassification,

DOI: 10.54254/2753-8818/2026.37886

their coverage was evaluated against the true gaps as a diagnostic; conventional likelihood intervals were not assumed valid for the misspecified physical model.

For each gap, bias was the mean estimation error, root mean square error (RMSE) was the square root of the mean squared error, and coverage was the proportion of intervals containing the generating value. Monte Carlo standard errors for bias were calculated as the standard deviation of estimates divided by $\sqrt{200}$. Coverage near 95% has a Monte Carlo standard error of approximately 1.54 percentage points with 200 repetitions [5].

Probability predictions were evaluated at the seven arithmetic midpoints between adjacent training temperatures. The mean Kullback–Leibler divergence from the clean generating distribution to the fitted distribution measured excess expected logarithmic loss, consistent with proper probabilistic scoring [8]. Thus, evaluation assessed recovery of physical populations, including when training labels were corrupted. It tested interpolation at unobserved temperatures, not extrapolation or laboratory validity.

4. Results

4.1 Temperature range and physical constraints

With clean labels and $n = 500$, zero-intercept constraints reduced upper-gap RMSE in every temperature range (Table 1). In the intermediate range, RMSE fell from 0.222 for unrestricted MNL to 0.061, a reduction of approximately 73%. However, temperature rankings depended on the estimand: the constrained lower-gap RMSE was smallest in the low range (0.025), whereas upper-gap RMSE was smallest in the intermediate range (0.061).

Table 1. Clean-data energy-gap RMSE by temperature range

Temperature range	Gap 1 RMSEConstrained	Gap 2 RMSEConstrained	Gap 2 RMSEUnrestricted
Low	0.025	0.075	0.299
Intermediate	0.039	0.061	0.222
High	0.235	0.220	0.809
Broad	0.042	0.099	0.136

$n = 500$ per temperature; 200 repetitions per condition. Gaps are in units of E_{ref}. Source: present simulations.

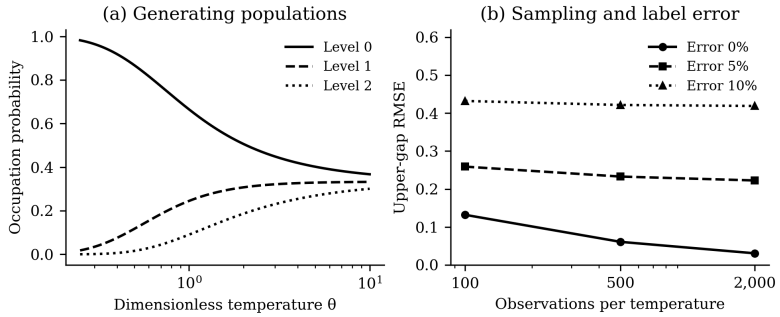

Figure 1. (a) Theoretical populations. (b) Upper-gap RMSE for constrained MNL in the intermediate range, averaged over 200 repetitions. Label error creates a persistent error floor. Source: present calculations and simulations

DOI: 10.54254/2753-8818/2026.37886

The broad range gave the lowest upper-gap RMSE for unrestricted MNL (0.136), indicating that temperature coverage was especially helpful when intercepts were unknown. High temperatures produced substantial gap uncertainty despite small predictive divergence: constrained RMSE values were 0.235 and 0.220, but mean divergence was only 0.000254 nats. Thus, a close match to populations need not imply precise energy recovery.

4.2 Sample size and uncertainty

In the intermediate clean-data condition, increasing n from 100 to 500 and 2,000 reduced constrained upper-gap RMSE from 0.132 to 0.061 and 0.031, approximately following inverse-square-root scaling (Figure 1). Corresponding lower-gap RMSE values were 0.087, 0.039 and 0.021. At $n = 2,000$, the mean estimated gaps were 0.999 and 2.001; Monte Carlo standard errors of their biases were 0.0015 and 0.0022. Upper-gap coverage was 96.0%, 93.5% and 94.0% across the three sample sizes. Coverage was not uniformly nominal across designs: the broad-range constrained upper-gap coverage at $n = 500$ was 90.0%, warranting caution about finite-sample Wald inference.

4.3 Misclassification and persistent bias

Misclassification systematically lowered estimated gaps. In the intermediate range at $n = 2,000$, the constrained upper-gap mean decreased from 2.001 with clean labels to 1.779 at 5% error and 1.582 at 10% error (Table 2). The latter is an approximately 21% underestimate. Upper-gap coverage fell from 94.0% to zero at both nonzero error rates.

Table 2. Misclassification effects for constrained MNL

Label error	Mean gap 1	Mean gap 2	Gap 2coverage	KL (nats)
0%	0.999	2.001	94.0%	0.000062
5%	0.931	1.779	0.0%	0.001817
10%	0.864	1.582	0.0%	0.006636

Intermediate temperatures; $n = 2,000$ per temperature; 200 repetitions. True gaps: 1 and 2. KL compares fitted and clean probabilities. Source: present simulations.

Additional observations did not remove this bias. With 10% error, the constrained upper-gap bias was -0.414 at $n = 100$ and -0.418 at $n = 2,000$; RMSE remained near 0.42. Unrestricted intercepts did not solve the problem: at $n = 2,000$ and 10% error, the unrestricted upper-gap mean was 1.398 and RMSE was 0.610. Nevertheless, its predictive divergence (0.006597 nats) was slightly smaller than the constrained value (0.006636), illustrating how similar population predictions can conceal different physical parameter errors.

5. Discussion

5.1 Why temperature selection matters

The simulations support the information-based interpretation in Equation 4. Low-temperature observations strongly favour the ground state, limiting evidence about the upper level. High-temperature observations populate all levels, but their populations change weakly with the energy gaps. Consequently, balanced class frequencies alone do not guarantee informative physical

DOI: 10.54254/2753-8818/2026.37886

measurements. The intermediate range offered the lowest upper-gap RMSE among the four tested ranges for the constrained estimator at $n = 500$; this is a comparison of specified designs, not a globally optimal temperature prescription.

The results are related to thermometric work demonstrating the importance of energy structure and measurement design [3, 4], but the estimands differ. An optimal design for measuring temperature cannot simply be assumed optimal for recovering two independent gaps. Moreover, the free-intercept model benefits from temperature variation because slopes and offsets must be separated. A broad range can therefore serve model diagnosis even when it does not minimise the constrained estimator's variance.

5.2 Physical knowledge and misspecification

Fixing the intercepts removes parameters that have known values under the nondegenerate canonical model. Its improved clean-data precision is therefore conditional on correct physical assumptions. Degenerate levels, temperature-dependent detection efficiencies or unmodelled states would require a different observation model. A nonzero fitted intercept may indicate such effects, but it cannot by itself establish a degeneracy ratio when detection is imperfect.

Misclassification makes the distinction between statistical fit and physical interpretation especially important. White's theory explains why maximum likelihood under misspecification can converge to a limiting parameter that differs from the physical truth [9]. Here, the population floor in Equation 5 reduces observed contrasts between levels. Increasing n then sharpens estimates around distorted gaps rather than correcting the distortion. Narrow intervals can consequently accompany poor coverage.

A calibrated confusion matrix could be incorporated directly into the likelihood through the mapping from physical to recorded probabilities. That extension was not fitted in this study, and its success should not be presumed when error rates are themselves unknown. Robust standard errors alone would address neither incorrect mean probabilities nor the resulting energy-gap bias.

5.3 Scope and limitations

The data were generated from an ideal equilibrium model. Recovering its parameters under matching assumptions checks computational estimation, not the validity of statistical mechanics. The design excludes interactions, quantum coherence, incomplete thermalisation and temporal dependence. Independent counts correspond to independent preparations or sufficiently separated measurements; correlated experimental observations would require additional modelling.

Temperature was treated as exact. An unknown multiplicative temperature calibration error is confounded with the overall energy scale, while random temperature error alters the predictor. The two assumed gaps and selected ranges also limit generalisation. Equal spacing was used only in data generation, but alternative spectra could change which design performs best. Finally, 200 repetitions provide moderate precision for comparisons and are insufficient for detailed analysis of rare numerical failures. Experimental application would require independent temperature calibration, validated state detection and explicit checks of thermal equilibrium.

6. Conclusion

Multinomial logistic regression provides an interpretable approach to estimating energy gaps from equilibrium state populations at calibrated temperatures. In the simulated three-level system,

DOI: 10.54254/2753-8818/2026.37886

imposing physically justified zero-intercept constraints improved estimation precision, while the most informative temperature range depended on the energy gap and model specification. Increasing sample size reduced random uncertainty but did not eliminate the systematic bias introduced by state misclassification. These findings demonstrate that accurate population predictions do not necessarily imply reliable recovery of physical parameters: both temperature selection and the measurement model must support the intended inference.

Future research could extend this framework to systems with additional energy levels, unequal energy spacing and known degeneracies, examining how spectral complexity affects parameter identifiability and sampling requirements. Incorporating a calibrated state-confusion matrix into the likelihood would allow measurement errors to be modelled explicitly, while accounting for temperature uncertainty would make the analysis more applicable to experimental settings. Adaptive sampling could also be investigated, with subsequent measurement temperatures selected using preliminary gap estimates and their uncertainty. Ultimately, application to independently measured state populations would help assess the framework's practical validity; extensions to interacting or nonequilibrium systems would require corresponding changes to the underlying physical probability model.

References

[1] McFadden, D. (1974). Conditional logit analysis of qualitative choice behavior. In P. Zarembka (Ed.), *Frontiers in econometrics* (pp. 105–142). Academic Press. https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf

[2] Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review, 106*(4), 620–630. https://doi.org/10.1103/PhysRev.106.620

[3] Correa, L. A., Mehboudi, M., Adesso, G., & Sanpera, A. (2015). Individual quantum probes for optimal thermometry. *Physical Review Letters, 114*(22), 220405. https://doi.org/10.1103/PhysRevLett.114.220405

[4] Hovhannisyan, K. V., Jørgensen, M. R., Landi, G. T., Alhambra, A. M., Brask, J. B., & Perarnau-Llobet, M. (2021). Optimal quantum thermometry with coarse-grained measurements. *PRX Quantum, 2*(2), 020322. https://doi.org/10.1103/PRXQuantum.2.020322

[5] Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine, 38*(11), 2074–2102. https://doi.org/10.1002/sim.8086

[6] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods, 17*(3), 261–272. doi:10.1038/s41592-019-0686-2.

[7] Albert, A., & Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika, 71*(1), 1–10. doi:10.1093/biomet/71.1.1.

[8] Gneiting, T., & Raftery, A. E. (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association, 102*(477), 359–378. doi:10.1198/016214506000001437.

[9] White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50*(1), 1–25. doi:10.2307/1912526.