Robust Design and Misspecification Detection for Count Models

Jordan Slessor[1], Rui Hu[12], and Linglong Kong[1]

[1]*Department of Mathematical and Statistical Sciences, University of Alberta*

[2]*Department of Mathematics and Statistics, MacEwan University*

Supplementary Material

These supplementary materials contain additional simulation results, a detailed description of the simulated annealing algorithm, and the full proofs of Theorems 1, 2, and 3. The extra simulations complement the main text by exploring scenarios with cubic misspecification, link function misspecification, high-dimensional design settings, and assessments of the variability of the simulated annealing algorithm. Results for the negative binomial model, omitted from the main text for space, are also included.

S1 Simulations

This section provides additional simulation results to support the findings presented in the main text. We include designs and performance summaries under cubic and link misspecification, as well as in a multivariate setting with a large number of design points. We also assess the variability of the simulated annealing algorithm across multiple runs. Finally, we present the negative binomial results for the primary experiments shown in the main text (no misspecification, and quadratic misspecification).

Unless otherwise stated, Gain and Premium are computed with respect to the design criterion used for optimization (e.g., $\mathcal{L}_{\text{full}}$ or $\mathcal{L}_{\text{ave}}$) and should not be interpreted as percentage changes in the Loss values reported in the Loss ($\mathcal{L}$) column. In particular, the $\mathcal{L}_{\text{ave}}$ criterion's theoretical loss scales linearly with its admissible-class radius ρ, so it is not on the same numerical scale as the classical or $\mathcal{L}_{\text{full}}$ criteria; it should be interpreted only relative to other $\mathcal{L}_{\text{ave}}$-criterion values.

S1.1 Negative Binomial: Simulation 1 (No Misspecification)

Figure 1 displays the design weight allocations under the negative binomial model for Simulation 1 (no misspecification). The figure complements the loss, Gain, and Premium results already reported for this model in Table 1 of the main text.

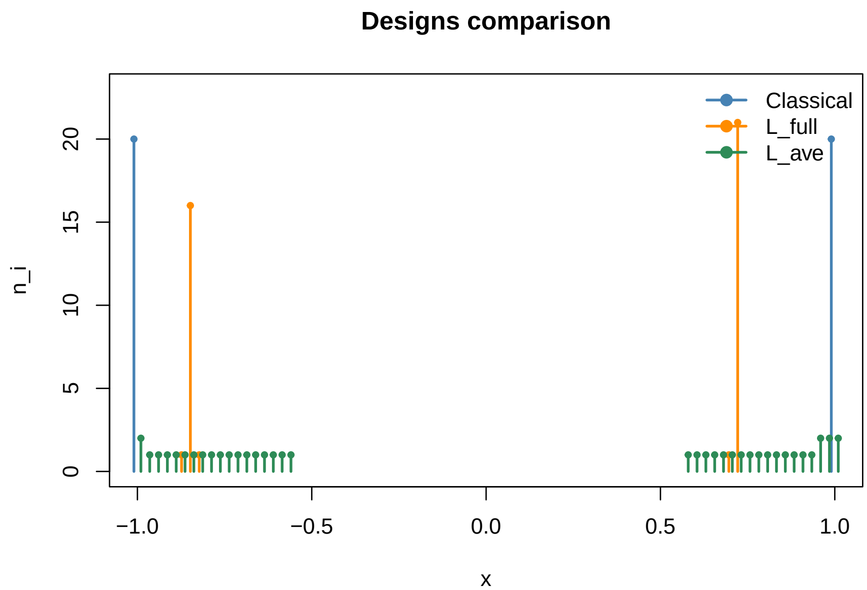

Figure 1: Design weight allocations for the negative binomial model under $\mathcal{L}_{\text{full}}$, $\mathcal{L}_{\text{ave}}$, and classical criteria (Simulation 1: No Misspecification).

S1.2 Negative Binomial: Simulation 2 (Quadratic Misspecification)

Figure 2 shows the estimated contamination and loess smoothed estimate that was used in the construction of the robust designs, complementing the results already reported in Table 2 of the main text.

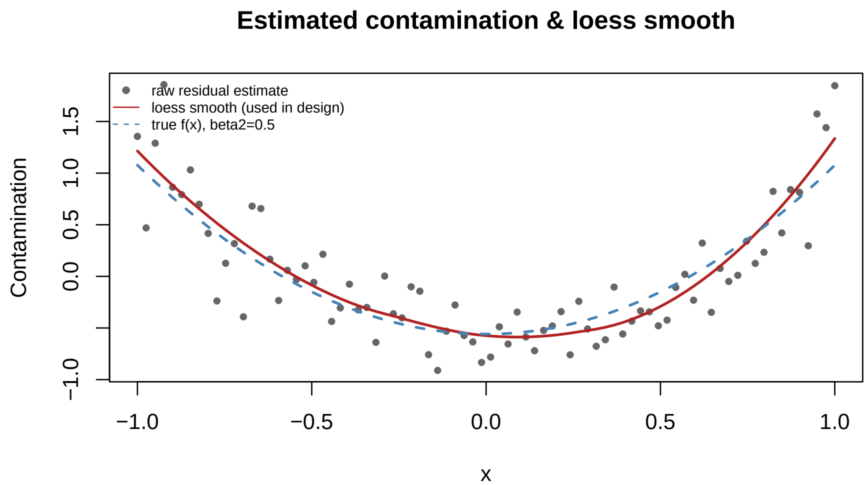

Figure 2: Estimated and loess smoothed contamination functions for negative binomial model (Simulation 2: Quadratic Misspecification).

S1.3 Simulation 4: Cubic misspecification

This simulation evaluates the impact of a third-order misspecification in the response model, using the same baseline configuration as the main text ($N = 80$ candidate points in $[-1, 1]$, $n = 40$, $\boldsymbol{z}(x) = (1, x)^\top$). The cubic contamination function is constructed as $f(x) = \beta_2\, g(x)$, where $g(x) = [x^3 - a_1 x - a_0]/\sqrt{N^{-1}\sum_i [x_i^3 - a_1 x_i - a_0]^2}$, with a_0, a_1 chosen by ordinary least squares so that $g(x)$ has zero sample mean and is uncorrelated with x; this enforces the same orthogonality and RMS-normalization conventions used for the quadratic contamination in the main text. As in Simulation 2, $\beta_2 = 0.5$ and $\boldsymbol{\beta} = (1/2, 1/2)^\top$. Design performance is evaluated using the same two-stage (for $\mathcal{L}_{\text{full}}$) / single-stage (for classical and $\mathcal{L}_{\text{ave}}$) simulated AMSE procedure described in the main text.

The XICOR test detects significant dependence under the Poisson model ($\xi = 0.740$, $Z = 10.459$, $p < 0.001$), but the negative binomial case also shows clear evidence of misspecification ($\xi = 0.492$, $Z = 6.953$, $p < 0.001$).

Table 1 summarizes the simulated AMSE, Gain, and Premium. As under quadratic misspecification, the classical design's realized loss deteriorates substantially (to 159.69 for Poisson and 168.13 for NB), while both robust designs cut this by more than half. The oracle Gain values confirm genuine protection against the true cubic contamination: 89.0% (Poisson, $\mathcal{L}_{\text{full}}$) and 98.7% (NB, $\mathcal{L}_{\text{full}}$), broadly comparable in magnitude to the quadratic-misspecification results in the main text, which is expected since the contamination here uses the same $\beta_2 = 0.5$ magnitude under the same RMS-normalization convention. The estimated contamination functions and design weights for the Poisson and Negative Binomial model are shown in Figures 3, and Figure 4, respectively.

S1.4 Simulation 5: Link misspecification

This simulation examines the effect of distributional (variance-function) misspecification in isolation from mean-structure misspecification. The

Table 1: Simulated AMSE, Gain, and Premium for Simulation 4 (Cubic Misspecification). Simulation standard errors (SE) are in parentheses. Gain (Est.) is evaluated at $\hat{f}$ (for $\mathcal{L}_{\text{full}}$) or $\rho = 10$ (for $\mathcal{L}_{\text{ave}}$); Gain (Oracle) is evaluated at the true contamination, $\beta_2 = 0.5$.

Design	Loss ($\mathcal{L}$)	Gain Est. (%)	Gain Oracle (%)	Premium (%)
Poisson				
Classical	159.6905 (1.6086)	–	–	–
Robust ($\mathcal{L}_{\text{full}}$)	67.8104 (0.8436)	87.2	89.0	17.7
Robust ($\mathcal{L}_{\text{ave}}$)	59.7297 (0.1485)	37.1	87.4	16.8
Negative Binomial				
Classical	168.1308 (4.6163)	–	–	–
Robust ($\mathcal{L}_{\text{full}}$)	71.8524 (1.7893)	99.1	98.7	21.7
Robust ($\mathcal{L}_{\text{ave}}$)	78.3362 (1.2678)	32.6	90.3	11.4

true model is Poisson with $\boldsymbol{z}(x) = (1, x)^\top$ over $N = 80$ points in $[-1, 1]$, $\mu(x) = \exp(\eta(x))$, $\eta(x) = \boldsymbol{z}(x)^\top\boldsymbol{\beta}$, $\boldsymbol{\beta} = (0.5, 0.5)^\top$, and $f(x) \equiv 0$ (no mean-structure misspecification). The experimenter, however, fits a negative binomial (NB2) model to a single Poisson pilot draw ($N = 80$ points, one observation each) and proceeds with design optimization using that fitted, misspecified NB2 model: both $\hat{\boldsymbol{\beta}}$ and $\hat{\theta}$ from this one fit are used directly as the working parameters for every design (classical, $\mathcal{L}_{\text{full}}$, $\mathcal{L}_{\text{ave}}$) and every loss computed below, since the fitted model itself stands in as

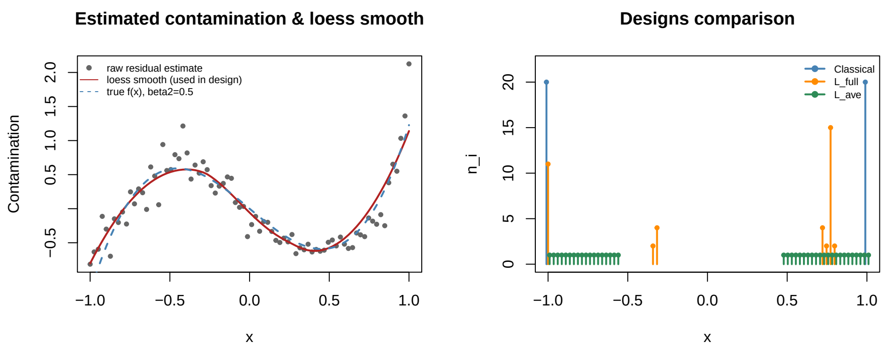

Figure 3: Estimated contamination function and design weight allocations for the Poisson model under $\mathcal{L}_{\text{full}}$, $\mathcal{L}_{\text{ave}}$, and classical criteria (Simulation 4: Cubic Misspecification).

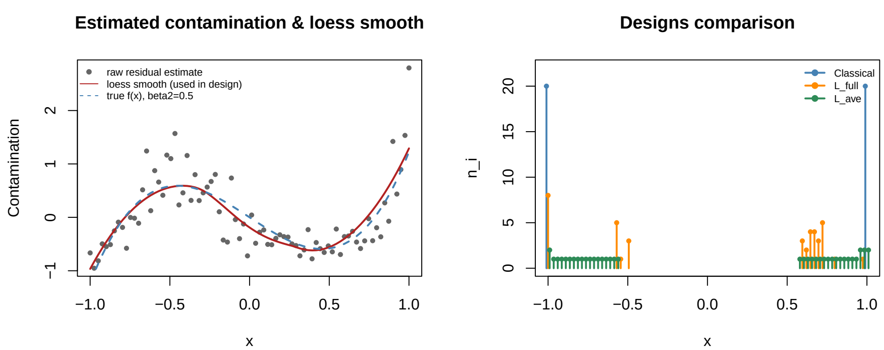

Figure 4: Estimated contamination function and design weight allocations for the negative binomial model under $\mathcal{L}_{\text{full}}$, $\mathcal{L}_{\text{ave}}$, and classical criteria (Simulation 4: Cubic Misspecification).

the assumed model. We note that with only one observation per design point, $\hat{\theta}$ is estimated close to the Poisson/NB2 boundary and can be highly variable from draw to draw; the Monte Carlo procedure used to evaluate

$\mathcal{L}_{\text{full}}$ integrates over this instability by drawing a fresh fit on every replicate.

The XICOR diagnostic, applied to residuals from the fitted (misspecified) model, gives $\xi = 0.059$, $Z = 0.828$, $p = 0.407$, showing no evidence of functional dependence – consistent with the fact that the misspecification here lies entirely in the assumed response family rather than the mean structure, which XICOR is not designed to detect.

Table 2 reports the simulated AMSE, Gain, and Premium, evaluated under the (misspecified) NB2 working model. The $\mathcal{L}_{\text{full}}$ design reduces loss modestly relative to the classical design (a 19.7% gain at a 6.2% premium), and $\mathcal{L}_{\text{ave}}$ achieves a larger 35.8% gain at a 15.9% premium. No oracle comparison is reported for this simulation, since the contamination here is distributional rather than a mean-structure function with a well-defined oracle counterpart.

Table 2: Simulated AMSE, Gain, and Premium for Simulation 5 (Link Misspecification: Poisson true, NB2 fit). Simulation standard errors (SE) are in parentheses. Gain and Premium are evaluated under the fitted NB2 model.

Design	Loss ($\mathcal{L}$)	Gain (%)	Premium (%)
Poisson (true), NB2 (fit)			
Classical	2.3787 (0.1136)	–	–
Robust ($\mathcal{L}_{\text{full}}$)	2.8193 (0.5167)	19.7	6.2
Robust ($\mathcal{L}_{\text{ave}}$)	2.8826 (0.1312)	35.8	15.9

To isolate the cost of adopting the misspecified NB2 model, Table 3 reports each design's loss when evaluated under the true Poisson model, alongside the benchmark loss of a classical design built with the correct Poisson model and true $\boldsymbol{\beta}$. All three NB2-derived designs incur a real cost relative to this benchmark, with the classical NB2-derived design coming closest (2.4507 versus the benchmark 2.3869) and $\mathcal{L}_{\text{ave}}$ farthest (2.7987). This indicates that, in this setting, the benefits of robustifying against an (apparently absent) mean-structure misspecification do not offset the cost of the underlying distributional misspecification when judged against the true data-generating model.

Table 3: Poisson-truth diagnostic for Simulation 5: loss of each NB2-derived design when evaluated under the true Poisson model, compared to the best-achievable Poisson-true benchmark.

Design	Loss under true Poisson
Classical (NB2-derived)	2.4507
Robust ($\mathcal{L}_{\text{full}}$, NB2-derived)	2.6084
Robust ($\mathcal{L}_{\text{ave}}$, NB2-derived)	2.7987
Classical (Poisson-true benchmark)	2.3869

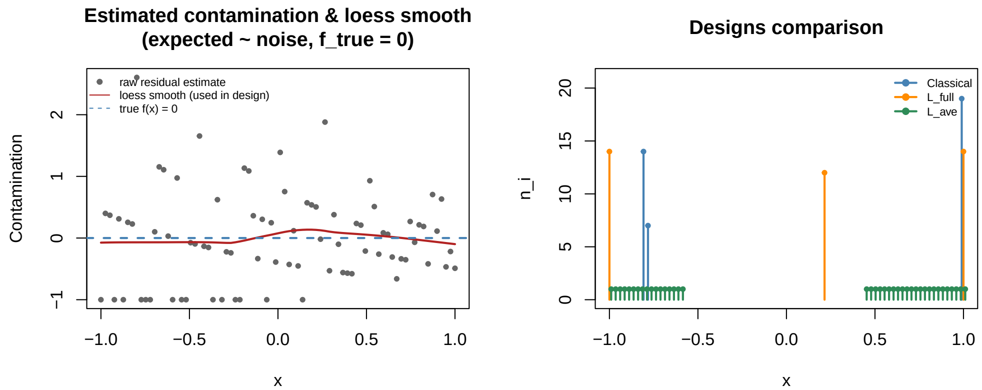

Figure 5: Estimated contamination function and design weight allocations under link misspecification (Poisson true, NB2 fit).

S1.5 Simulation 6: Multivariate, large N

This simulation investigates the performance of robust designs in a multivariate setting with a larger design space and structured misspecification, to address concerns about the scalability and practical relevance of the proposed methods beyond univariate cases.

We consider a bivariate design space with $N = 225$ candidate points formed by a 15×15 regular grid over $[-1, 1]^2$. The covariate vector is $\boldsymbol{z}(x) = (1, x_1, x_2)^\top$, and the true linear predictor is $\eta(x) = \boldsymbol{z}(x)^\top \boldsymbol{\beta} + f(x_1, x_2)$ with $\boldsymbol{\beta} = (0.5, 0.5, -0.5)$. The misspecification function is $f(x_1, x_2) = \tau\, g(x_1, x_2)$ with $\tau = \sqrt{0.25} = 0.5$, where $g(x_1, x_2)$ is the interaction term $x_1 x_2$ projected orthogonal to $\{1, x_1, x_2\}$ via ordinary least squares and then RMS-normalized, so that $N^{-1}\sum_i f(x_i)^2 = \tau^2 = 0.25$ exactly, satisfying the bounded bias and identifiability conditions. Responses are generated as $Y_i \sim \text{Poisson}(\mu_i)$, $\mu_i = \exp(\eta_i)$. We compare classical ($f = 0$), $\mathcal{L}_{\text{full}}$ (contamination estimated from $N = 225$ pilot points with 5 replicates each), and $\mathcal{L}_{\text{ave}}$ designs, allocating the baseline $n = 40$ total observations across the $N = 225$ candidate points.

Because XICOR is defined for a scalar covariate, we test residuals against x_1 and x_2 separately (Bonferroni-corrected at $\alpha = 0.0167$ per component) and against the interaction term $x_1 x_2$, the most powerful direction for detecting this particular misspecification. No dependence is detected against either marginal covariate ($\xi = 0.030$, $p = 0.482$ for x_1; $\xi = 0.010$, $p = 0.805$ for x_2), but the test strongly rejects independence against the interaction term ($\xi = 0.373$, $p < 0.001$), correctly flagging the structured (interaction) misspecification at family-wise $\alpha = 0.05$.

Given the computational cost of re-optimizing a design at $N = 225$ candidate points – and as with the real-data earthworm example in the main text – we report theoretical losses only; no simulated AMSE is computed for this setting. Table 4 summarizes the results, together with the size of each design's support. The classical design concentrates all $n = 40$ observations on just 4 of the 225 candidate points, while $\mathcal{L}_{\text{full}}$ spreads its support over 11 points and $\mathcal{L}_{\text{ave}}$ spreads it over all 40 (one observation per allocated point). As flagged in the introduction to this section, the robust designs' theoretical loss values are evaluated under their own (penalized) criteria and are not on the same scale as the classical design's $f = 0$ loss; the Gain and Premium columns are the appropriate comparison. By that measure, $\mathcal{L}_{\text{full}}$ achieves a 98.6% estimated gain (92.7% at the oracle truth $\tau = 0.5$) at a 29.6% premium, and $\mathcal{L}_{\text{ave}}$ achieves a 30.7% estimated gain (77.4% oracle) at a more modest 17.8% premium – confirming that the robust design framework extends to multivariate settings and remains effective under structured interaction misspecification.

Table 4: Theoretical loss, support size, Gain, and Premium for Simulation 6 (Multivariate Interaction Misspecification). Gain (Est.) is evaluated at $\hat{f}$ (for $\mathcal{L}_{\text{full}}$) or $\rho = 10$ (for $\mathcal{L}_{\text{ave}}$); Gain (Oracle) is evaluated at the true contamination, $\tau = 0.50$.

Design	Supp.	Loss ($\mathcal{L}$)	Gain Est. (%)	Gain Oracle (%)	Premium (%)
Classical	4	3.2513	–	–	–
Robust ($\mathcal{L}_{\text{full}}$)	11	42.5017	98.6	92.7	29.6
Robust ($\mathcal{L}_{\text{ave}}$)	40	43.6348	30.7	77.4	17.8

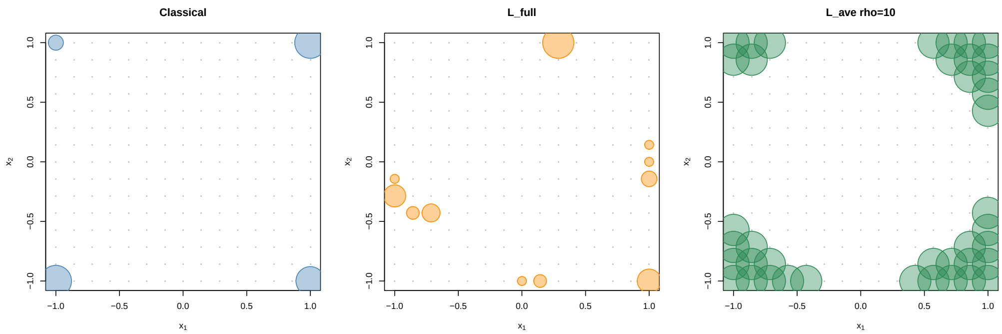

Figure 6: Design weight allocations.

S1.6 Simulation 7: Multiple Annealing Runs

This simulation examines the run-to-run variability of the simulated annealing algorithm. We use the Poisson model with $\boldsymbol{z}(x) = (1, x)^\top$ over $N = 40$ evenly spaced points in $[-1, 1]$, $n = 40$, $\boldsymbol{\beta} = (0.5, 0.5)^\top$, and the same quadratic misspecification as Simulation 2 of the main text ($f_{\text{true}}(x) = 0.5\, g(x)$, with $g(x)$ the RMS-normalized quadratic contamination). With $N = n = 40$, the symmetric initial design places exactly one observation at every candidate point.

For each of $\mathcal{L}_{\text{full}}$ and $\mathcal{L}_{\text{ave}}$, we run the simulated annealing algorithm 100 times from this same symmetric starting point and record, at each design point, the mean and standard deviation of the resulting allocation across runs; $\mathcal{L}_{\text{full}}$ uses a single, fixed pilot-estimated objective across all 100 runs, so that all variability reflects the optimizer's own randomness rather than pilot-data variability. The mean designs are shown in Figure 7, with error bars giving one standard deviation at each point.

Unlike $\mathcal{L}_{\text{ave}}$, which converged to the exact same design in all 100 runs (loss SD = 0.0000; weight SD = 0 at every point), $\mathcal{L}_{\text{full}}$ shows real run-to-run variability: the realized loss ranged from 4.4038 to 4.5729 across runs (SD = 0.0239 on a mean of 4.4464), and individual design weights varied by as much as an SD of 4.19 at the points carrying the most mass (max mean weight 9.08). The effective support size also fluctuated somewhat more for $\mathcal{L}_{\text{full}}$ (mean 23.0 points) than for $\mathcal{L}_{\text{ave}}$ (mean 26.0 points, with zero variability). Table 5 summarizes these aggregate statistics.

Table 5: Summary of variability across 100 simulated annealing runs for Simulation 7, by loss criterion.

Metric	$\mathcal{L}_{\text{full}}$	$\mathcal{L}_{\text{ave}}$
Mean loss	4.4464	4.7663
SD loss	0.0239	0.0000
Min loss	4.4038	4.7663
Max loss	4.5729	4.7663
Max mean weight	9.080	3.000
Max SD weight	4.189	0.000
Effective support (mean)	23.0	26.0

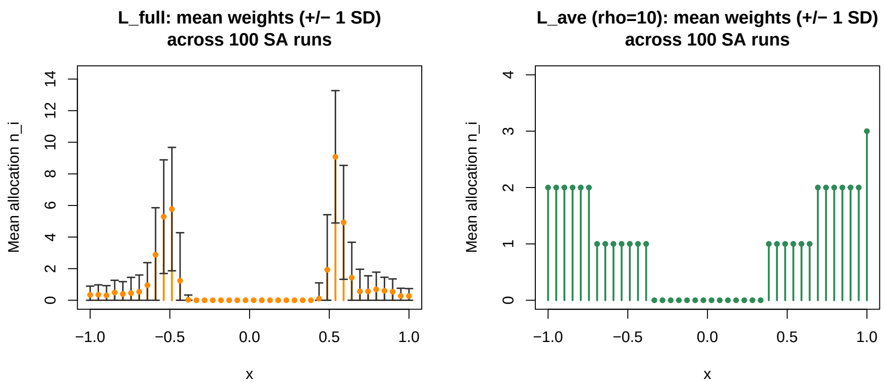

Figure 7: Mean design weight allocations across 100 annealing runs, with error bars showing ± 1 SD at each candidate point, for $\mathcal{L}_{\text{full}}$ and $\mathcal{L}_{\text{ave}}$.

While $\mathcal{L}_{\text{ave}}$'s objective surface appears to have a single, easily identified optimum, the $\mathcal{L}_{\text{full}}$ objective shows meaningfully different optima (or different points reached within the SA budget) across runs from the same starting design. In practice, this suggests that $\mathcal{L}_{\text{full}}$-based designs should be selected as the best of several independent annealing runs (e.g., lowest realized loss) rather than relying on a single run, whereas a single run appears adequate for $\mathcal{L}_{\text{ave}}$.

S2 Construction of Robust Designs via Simulated Annealing

Because the derived robust loss functions operate over discrete parameter spaces and are highly non-convex, traditional gradient-based optimization techniques are inapplicable. We therefore implement a simulated annealing algorithm to optimize the design weights. Initializing from a uniformly spaced grid, the algorithm proposes candidate designs utilizing the perturbation method of Fang and Wiens (2000), which iteratively shifts allocated design points to neighboring locations while preserving design symmetry.

Let $\mathcal{L}$ and $\mathcal{L}'$ denote the loss evaluated at the current and proposed designs, respectively, and define the change in energy as $\Delta E = \mathcal{L}' - \mathcal{L}$. Proposed configurations that reduce the loss ($\Delta E < 0$) are accepted deterministically. To escape local minima, suboptimal moves ($\Delta E \geq 0$) are accepted with probability $\exp(-\Delta E/T)$, permitting broad exploration when the system temperature (T) is high. The algorithm utilizes a logarithmic cooling schedule: every 1,000 iterations, the temperature is updated as $T = T_{init}/\log(e + 0.0001 \times i)$, where i is the current iteration and T_{init} is the initial starting temperature. The probability of accepting inferior designs smoothly decays until the maximum iteration threshold is reached, ensuring convergence to a robust optimal allocation.

S3 Proofs

In this section we prove Theorems 1, 2, and 3 from the original manuscript.

S3.1 Proof of Theorem 1

Proof of Theorem 1. The proof follows from that of Adewale and Wiens (2009). Under the conditions of Fahrmeir (1990), including bounded misspecification f, full rank of the design matrix, and standard regularity conditions on the log-likelihood, the maximum likelihood estimate $\hat{\boldsymbol{\beta}}$ exists, is consistent for the target parameter $\boldsymbol{\beta}_0$, and satisfies $\partial l(\hat{\boldsymbol{\beta}})/\partial\boldsymbol{\beta} = o_p(n^{-1/2})$. These assumptions, originally used in the misspecified logistic model setting of Adewale and Wiens (2009), apply similarly in the count model context considered here.

The log-likelihood l, the score function, and the negative second derivative according to the assumed model are

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left\{ n_i \left[y_i \eta_i - e^{\eta_i} - \log(y_i!\,) \right] \right\},$$

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{N} n_i (y_i - e^{\eta_i}) \boldsymbol{z}(\boldsymbol{x}_i),$$

and

$$-\frac{\partial^2 l(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = \sum_{i=1}^{N} n_i e^{\eta_i} \boldsymbol{z}(\boldsymbol{x}_i)\boldsymbol{z}^\top(\boldsymbol{x}_i),$$

respectively. Now a Taylor expansion of $\partial l(\boldsymbol{\beta})/\partial\beta_j$ around $\boldsymbol{\beta}_0$ yields

$$\begin{aligned}\frac{\partial l(\boldsymbol{\beta})}{\partial\beta_j} &= \frac{\partial l(\boldsymbol{\beta}_0)}{\partial\beta_j} + \sum_{k=1}^{p}(\beta_k - \beta_{0,k})\frac{\partial^2 l(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_k} \\ &\quad + \frac{1}{2}\sum_{k=1}^{p}\sum_{l=1}^{p}(\beta_k - \beta_{0,k})(\beta_l - \beta_{0,l})\frac{\partial^3 l(\boldsymbol{\beta}_*)}{\partial\beta_j\partial\beta_k\partial\beta_l},\end{aligned}$$

where β_j and $\beta_{0,j}$ are the jth terms of $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$, respectively, and $\boldsymbol{\beta}_*$ is a point on the line segment connecting $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_0$. Now replace $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}}$ and multiply everything by $n^{-1/2}$ to obtain

$$\begin{aligned}-\frac{1}{\sqrt{n}}\frac{\partial l(\boldsymbol{\beta}_0)}{\partial\beta_j} &= \frac{1}{\sqrt{n}}\sum_{k=1}^{p}(\hat{\beta}_k - \beta_{0,k})\frac{\partial^2 l(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_k} \\ &\quad + \frac{1}{2\sqrt{n}}\sum_{k=1}^{p}\sum_{l=1}^{p}(\hat{\beta}_k - \beta_{0,k})(\hat{\beta}_l - \beta_{0,l})\frac{\partial^3 l(\boldsymbol{\beta}_*)}{\partial\beta_j\partial\beta_k\partial\beta_l}.\end{aligned}$$

Now rearrange to get

$$\begin{aligned}-\frac{1}{\sqrt{n}}\frac{\partial l(\boldsymbol{\beta}_0)}{\partial\beta_j} = \sqrt{n}\sum_{k=1}^{p}(\hat{\beta}_k - \beta_{0,k})\Bigg[&\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_k} \\ &+\frac{1}{2n}\sum_{l=1}^{p}(\hat{\beta}_l - \beta_{0,l})\frac{\partial^3 l(\boldsymbol{\beta}_*)}{\partial\beta_j\partial\beta_k\partial\beta_l}\Bigg].\end{aligned}$$

When the response follows a Poisson distribution, and when using a negative binomial response with $\mu_i = e^{\eta_i}$, $\partial^3 l(\boldsymbol{\beta}_*)/\partial\beta_j\partial\beta_k\partial\beta_l$ is bounded. Thus, since $\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}_0$,

$$\left[\frac{1}{n}\frac{\partial^2 l(\boldsymbol{\beta}_0)}{\partial\beta_j\partial\beta_k} + \frac{1}{2n}\sum_{l=1}^{p}(\hat{\beta}_l - \beta_{0,l})\frac{\partial^3 l(\boldsymbol{\beta}_*)}{\partial\beta_j\partial\beta_k\partial\beta_l}\right] + H_{jk} \xrightarrow{p} 0,$$

where H_{jk} is the (j,k)th element of $\boldsymbol{H}_n$. Now, since $\sum_{k=1}^{p} H_{jk}\sqrt{n}(\hat{\beta}_k - \beta_{0,k}) = (1/\sqrt{n})\partial l(\boldsymbol{\beta}_0)/\partial \beta_j$, solving for $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ yields

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) = \boldsymbol{H}_n^{-1}(1/\sqrt{n})\partial l(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}.$$

Thus, the limit distribution of $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is the same as the limit distribution of $\boldsymbol{H}_n^{-1}(1/\sqrt{n})\partial l(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}$. Then using the central limit theorem for independent but not identically distributed random variables, $(1/\sqrt{n})\partial l(\boldsymbol{\beta}_0)/\partial \boldsymbol{\beta}$ has a multivariate normal limit distribution with mean $(1/\sqrt{n})\sum_{i=1}^{N} n_i E[y_i - e^{\boldsymbol{z}(\boldsymbol{x}_i)\boldsymbol{\beta}_0}]\boldsymbol{z}(\boldsymbol{x}_i) = \sqrt{n}\boldsymbol{b}$ and covariance matrix $\tilde{\boldsymbol{H}}_n = \boldsymbol{Z}^\top \boldsymbol{P}\boldsymbol{W}_T\boldsymbol{Z}$. Finally, from this $\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)$ is asymptotically normal with mean $\sqrt{n}\boldsymbol{H}_n^{-1}\boldsymbol{b}$ and covariance $\boldsymbol{H}_n^{-1}\tilde{\boldsymbol{H}}_n\boldsymbol{H}_n^{-1}$. □

S3.2 Proof of Theorem 2

Proof of Theorem 2. This theorem and proof follows corollary 2 from Adewale and Wiens (2009). First decompose the AMSE as

$$\mathcal{L} = \frac{1}{nN}\sum_{i=1}^{N} var[\sqrt{n}\mu(\hat{\eta}_i)] + \frac{1}{nN}\sum_{i=1}^{N}\{E[\sqrt{n}\mu(\hat{\eta}_i)] - \sqrt{n}\mu(\eta_i + f(\boldsymbol{x}_i))\}^2.$$

By the Delta method, the first sum is

$$\begin{aligned}\frac{1}{nN}\sum_{i=1}^{N} var[\sqrt{n}\mu(\hat{\eta}_i)] &= \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{d\mu_i}{d\eta_i}\right)^2 var[\sqrt{n}\hat{\eta}_i] + o(n^{-1}) \\ &= \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{d\mu_i}{d\eta_i}\right)^2 \boldsymbol{z}^\top(\boldsymbol{x}_i)\boldsymbol{H}_n^{-1}\tilde{\boldsymbol{H}}_n\boldsymbol{H}_n^{-1}\boldsymbol{z}(\boldsymbol{x}_i) + o(n^{-1}) \\ &= \frac{1}{nN} tr[\boldsymbol{W}\boldsymbol{Z}\boldsymbol{H}_n^{-1}\tilde{\boldsymbol{H}}_n\boldsymbol{H}_n^{-1}\boldsymbol{Z}^\top\boldsymbol{W}] + o(n^{-1}).\end{aligned}$$

Now expand $\mu(\hat{\eta}_i)$ and $\mu(\eta_i + f(\boldsymbol{x}_i))$ around η_i to get

$$E[\sqrt{n}\mu(\hat{\eta}_i)] = \sqrt{n}\mu(\eta_i) + E\left[\sqrt{n}\frac{d\mu}{d\eta_i}(\hat{\eta}_i - \eta_i) + o(\sqrt{n}(\hat{\eta}_i - \eta_i))\right],$$

and

$$\sqrt{n}\mu(\eta_i + f(\boldsymbol{x}_i)) = \sqrt{n}\mu(\eta_i) + \sqrt{n}\frac{d\mu}{d\eta_i} f(\boldsymbol{x}_i) + o(\sqrt{n} f(\boldsymbol{x}_i)),$$

respectively. Then

$$E[\sqrt{n}\mu(\hat{\eta}_i)] = \sqrt{n}\mu(\eta_i) + \sqrt{n}\frac{d\mu}{d\eta_i}E[\hat{\eta}_i - \eta_i] + o(1),$$

and therefore the second sum in the decomposition of the AMSE is

$$\begin{aligned}&\frac{1}{nN}\sum_{i=1}^{N}\{E[\sqrt{n}\mu(\hat{\eta}_i)] - \sqrt{n}\mu(\eta_i + f(\boldsymbol{x}_i))\}^2 \\ &= \frac{1}{nN}\sum_{i=1}^{N}\left(\frac{d\mu}{d\eta_i}\right)^2 \{n \cdot bias^\top(\hat{\boldsymbol{\beta}})\boldsymbol{z}(\boldsymbol{x}_i)\boldsymbol{z}^\top(\boldsymbol{x}_i) bias(\hat{\boldsymbol{\beta}}) + nf^2(\boldsymbol{x}_i)\} + o(n^{-1}) \\ &= \frac{1}{nN}\{n\boldsymbol{b}^\top\boldsymbol{H}_n^{-1}\boldsymbol{Z}^\top\boldsymbol{W}^2\boldsymbol{Z}\boldsymbol{H}_n^{-1}\boldsymbol{b} - 2n\boldsymbol{f}^\top\boldsymbol{W}^2\boldsymbol{Z}\boldsymbol{H}_n^{-1}\boldsymbol{b} + n\boldsymbol{f}^\top\boldsymbol{W}^2\boldsymbol{f}\} + o(n^{-1}) \\ &= \frac{1}{N}||\boldsymbol{W}(\boldsymbol{Z}\boldsymbol{H}_n^{-1}\boldsymbol{b} - \boldsymbol{f})||^2 + o(n^{-1}).\end{aligned}$$

$\square$

S3.3 Proof of Theorem 3

$$\frac{1}{N}\sum_{i=1}^{N} f^2(\boldsymbol{x}_i) \leq \tau^2, \tag{S3.1}$$

where $\tau^2 = O(n^{-1})$, and

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{z}(\boldsymbol{x}_i) f(\boldsymbol{x}_i) = \boldsymbol{0}_{p\times 1}. \tag{S3.2}$$

Next, the singular value decomposition of the covariate matrix

$$\boldsymbol{Z} = \boldsymbol{U}_{N\times p}\boldsymbol{\Lambda}_{p\times p}\boldsymbol{V}_{p\times p}^{\top}, \tag{S3.3}$$

where $\boldsymbol{U}^{\top}\boldsymbol{U} = \boldsymbol{V}^{\top}\boldsymbol{V} = \boldsymbol{I}_p$ and $\boldsymbol{\Lambda}$ is diagonal and invertible, followed by the augmentation of $\boldsymbol{U}$, $\tilde{\boldsymbol{U}}_{N\times(N-p)}$, such that $[\boldsymbol{U}\vdots\tilde{\boldsymbol{U}}]_{N\times N}$ is orthogonal. Then by assumptions (S3.1) and (S3.2), there exists an $(N-p)\times 1$ vector $\boldsymbol{t}$, with $||\boldsymbol{t}|| \leq 1$ satisfying

$$\boldsymbol{f}(=\boldsymbol{f}_{\boldsymbol{t}}) = \tau\sqrt{N}\tilde{\boldsymbol{U}}\boldsymbol{t}. \tag{S3.4}$$

Proof of Theorem 3. In this proof we will use the identity $\int \boldsymbol{t}^{\top}\boldsymbol{t}p(\boldsymbol{t})d\boldsymbol{t} = (N-p)/(N-p+2)$, implying that

$$\int \boldsymbol{t}\boldsymbol{t}^{\top}p(\boldsymbol{t})d\boldsymbol{t} = \frac{1}{N-p+2}\boldsymbol{I}_{N-p}.$$

Now use (S3.3) and (S3.4) to write $\mathcal{L}_{\text{full}}$ in terms of $\boldsymbol{t}$ as

$$\begin{aligned}\mathcal{L}_{\text{full}}(\boldsymbol{P},\boldsymbol{f}) = \frac{1}{nN}\Big\{&tr[(\boldsymbol{U}^{\top}\boldsymbol{P}\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{U}^{\top}\boldsymbol{P}\boldsymbol{W}_T(\boldsymbol{t})\boldsymbol{U}(\boldsymbol{U}^{\top}\boldsymbol{P}\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{U}^{\top}\boldsymbol{W}^2\boldsymbol{U}]\\ &+||\boldsymbol{W}(\boldsymbol{U}(\boldsymbol{U}^{\top}\boldsymbol{P}\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{U}^{\top}\boldsymbol{P}(\boldsymbol{\gamma}_T(\boldsymbol{t})-\boldsymbol{\gamma}) - \tau\sqrt{N}\tilde{\boldsymbol{U}}\boldsymbol{t})||^2\Big\}\end{aligned} \tag{S3.5}$$

Via another application of (S3.4), we expand $\boldsymbol{W}_T(\boldsymbol{t})$ elementwise as

$$\boldsymbol{W}_T(\boldsymbol{t}) = \boldsymbol{W} + diag(\dot{\boldsymbol{w}} \circ (\tilde{\boldsymbol{U}}\boldsymbol{t}))\tau\sqrt{N} + O(\tau^2),$$

where $\dot{\boldsymbol{w}} = (dw_i/d\eta_i)_{i=1}^N$ and $\circ$ denotes the Hadamard (elementwise) product. Now since $\tau^2 = O(n^{-1})$, $\int \boldsymbol{W}_T(\boldsymbol{t})p(\boldsymbol{t})d\boldsymbol{t} = \boldsymbol{W} + O(n^{-1})$, and so

$$\begin{aligned}&\int tr[(\boldsymbol{U}^\top \boldsymbol{P}\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{P}\boldsymbol{W}_T(\boldsymbol{t})\boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{P}\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{W}^2\boldsymbol{U}]p(\boldsymbol{t})d\boldsymbol{t}\\ &= tr[(\boldsymbol{U}^\top \boldsymbol{P}\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{W}^2\boldsymbol{U}] + O(n^{-1}).\end{aligned}$$

By a similar argument, $\boldsymbol{\gamma}_T(\boldsymbol{t}) - \boldsymbol{\gamma} = \tau\sqrt{N}\boldsymbol{W}\tilde{\boldsymbol{U}}\boldsymbol{t} + O(\tau^2)$, and so

$$\begin{aligned}&||\boldsymbol{W}(\boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{P}\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{P}(\boldsymbol{\gamma}_T(\boldsymbol{t}) - \boldsymbol{\gamma}) - \tau\sqrt{N}\tilde{\boldsymbol{U}}\boldsymbol{t})||^2\\ &= \tau^2 N||\boldsymbol{W}(\boldsymbol{R} - \boldsymbol{I})\tilde{\boldsymbol{U}}\boldsymbol{t}||^2 + O(n^{-1/2}),\end{aligned}$$

with

$$\begin{aligned}&\int ||\boldsymbol{W}(\boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{P}\boldsymbol{W}\boldsymbol{U})^{-1}\boldsymbol{U}^\top \boldsymbol{P}(\boldsymbol{\gamma}_T(\boldsymbol{t}) - \boldsymbol{\gamma}) - \tau\sqrt{N}\tilde{\boldsymbol{U}}\boldsymbol{t})||^2 p(\boldsymbol{t})d\boldsymbol{t}\\ &= \frac{\tau^2 N tr[\boldsymbol{W}(\boldsymbol{R} - \boldsymbol{I})\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^\top(\boldsymbol{R} - \boldsymbol{I})^\top \boldsymbol{W}]}{N - p + 2}.\end{aligned}$$

The result of the theorem then follows from the fact that $\tilde{\boldsymbol{U}}\tilde{\boldsymbol{U}}^\top = \boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top$ and $(\boldsymbol{R} - \boldsymbol{I})\boldsymbol{U} = \boldsymbol{0}$ alongside substituting these integrals into (S3.5). $\square$

Bibliography

Adewale, A. J. and D. P. Wiens (2009). Robust designs for misspecified logistic models. *Journal of Statistical Planning and Inference 139*(1), 3–15. Special Issue on Metaheuristics, Combinatorial Optimization and Design of Experiments.

Fahrmeir, L. (1990). Maximum likelihood estimation in misspecified generalized linear models. *Statistics 21*(4), 487–502.

Fang, Z. and D. P. Wiens (2000). Integer-valued, minimax robust designs for estimation and extrapolation in heteroscedastic, approximately linear models. *Journal of the American Statistical Association 95*(451), 807–818.