When Does Model Simplification Matter? Consequence Analysis for Weibull Series Systems

Alex Towell
Department of Computer Science
Southern Illinois University Edwardsville
lex@metafunctor.com

Abstract

When a reliability engineer simplifies a series system model by assuming all components share a common Weibull shape parameter, how much accuracy is lost—and can the data even tell the difference? We show that these two questions have a quantitatively aligned answer. The common-shape model's prediction bias grows super-linearly with shape heterogeneity, remaining below 1% in system MTTF through shape CV $\approx 15\%$, while the likelihood ratio test's power to detect the misspecification grows super-linearly. This creates a safe zone: the test lacks power precisely where the consequence of using the wrong model is negligible for engineering decisions. An adaptive LRT-based procedure exploits this alignment, achieving RMSE within 2.5% of the full model at $n \geq 500$ while selecting the simpler model over 90% of the time when appropriate. The common-shape model is the unique single-parameter reduction that preserves the Weibull property of the system lifetime, making it the natural parsimony boundary. Our analysis, conducted under realistic conditions of right-censored and masked failure data, provides practitioners with both the assurance that simplification is safe and the tools to decide when it is not.

1 Introduction

Consider a turbine engine whose operation depends on bearings, seals, and blades arranged in series—any component failure grounds the engine. Field maintenance logs record the failure time and a short list of suspect components (the candidate set), but exact root-cause analysis requires costly teardown. Warranty programs further right-censor units still in service. The engineer faces a fundamental modeling choice: a full model with separate Weibull shape and scale parameters for each of m components ($2m$ parameters), or a reduced model that assumes a common shape parameter across all components ($m + 1$ parameters).

The full model is more flexible but harder to estimate with limited masked data. The reduced model is more parsimonious and—uniquely among single-parameter reductions—renders the system lifetime itself Weibull [1], enabling closed-form reliability metrics. But is the simplification safe? The existing literature on this question focuses almost entirely on whether a likelihood ratio test can *detect* the difference between models [2]. This is the wrong question. The right question is: when the reduced model is wrong, does it produce predictions that are wrong enough to change an engineering decision?

This paper answers both questions and quantifies the alignment between them. The prediction bias of the common-shape model grows super-linearly with shape heterogeneity, while the LRT's power to detect that heterogeneity grows super-linearly. The consequence is negligible precisely

where the test lacks power, and the test has high power precisely where the consequence warrants switching models. An adaptive procedure exploits this alignment to automatically select the appropriate model from data.

1.1 Related Work

The statistical treatment of masked system failure data originates with Usher and Hodgson [3], who introduced maximum likelihood methods for component reliability estimation from masked system life-test data. Lin, Usher, and Guess [4] derived exact maximum likelihood estimators, while Lin, Usher, and Guess [5] developed Bayesian approaches. For Weibull components specifically, Usher [6] addressed reliability prediction with masked data, and Guess and Usher [7] proposed an iterative estimation approach. Sarhan [8, 9] extended the framework to various distributional assumptions, Tan [10] treated masked binomial system testing data, while Tan [11] contributed methods for exponential component reliability estimation. Guo, Niu, and Szidarovszky [12] studied component reliability estimation from incomplete system failure data, providing the baseline system configuration used in our simulations. Towell [2] developed a comprehensive likelihood model incorporating both right-censoring and candidate sets for Weibull series systems.

The closure property of the Weibull minimum under common shape is a classical result in reliability theory [1, 13, 14]. Craiu and Lee [15] addressed model selection for competing-risks models with and without masking, though not in the Weibull series system context. The model selection question—common shape versus heterogeneous shapes—has not, to our knowledge, been studied systematically for masked data, nor has the consequence of misspecification been quantified.

For general model selection methodology, we employ the likelihood ratio test [16] and compare with the Akaike Information Criterion [17] and Bayesian Information Criterion [18]. For a comprehensive treatment of information-theoretic model selection, see [19].

1.2 Contributions

1. **Bias-detectability alignment**: We quantify the alignment between the common-shape model's prediction bias and the LRT's detection power as functions of shape heterogeneity. MTTF bias remains below 1% through CV $\approx 15\%$ (super-linear growth), while the LRT achieves 80% power only at higher CVs (super-linear growth). The practical consequence: the wrong model choice is harmless exactly where it is likely.

2. **Consequence analysis**: We quantify the bias in system MTTF and reliability function predictions when the common-shape model is applied to systems with heterogeneous shapes. The 1% MTTF bias threshold is not crossed until CV $\approx 20\%$. Surprisingly, a bias-variance decomposition reveals that the full model has lower MTTF variance than the reduced model even when the reduced model is correctly specified, because the nonlinear MTTF functional amplifies the constrained estimator's variance more than the unconstrained estimator's.

3. **Adaptive model selection**: We evaluate an LRT-based procedure that exploits the alignment, achieving RMSE within 2.5% of the always-full strategy at $n \geq 500$ while selecting the simpler model $> 90\%$ of the time when the data support it.

The paper is organized as follows. Section 2 presents the model framework. Section 3 quantifies the consequences of model misspecification. Section 4 characterizes the LRT. Section 5 develops the adaptive procedure. Section 6 concludes.

2 Model Framework

2.1 Series System with Weibull Components

Consider a series system of m independent components, where the system fails when any component fails. Each component lifetime T_j follows a two-parameter Weibull distribution with shape $k_j > 0$ and scale $\lambda_j > 0$:

$$f_j(t; k_j, \lambda_j) = \frac{k_j}{\lambda_j}\left(\frac{t}{\lambda_j}\right)^{k_j-1} \exp\left\{-\left(\frac{t}{\lambda_j}\right)^{k_j}\right\}, \quad t > 0. \tag{1}$$

The system lifetime $T = \min\{T_1, \ldots, T_m\}$ has reliability function, hazard function, and density:

$$R(t; \boldsymbol{\theta}) = \exp\left\{-\sum_{j=1}^{m}\left(\frac{t}{\lambda_j}\right)^{k_j}\right\}, \tag{2}$$

$$h(t; \boldsymbol{\theta}) = \sum_{j=1}^{m}\frac{k_j}{\lambda_j}\left(\frac{t}{\lambda_j}\right)^{k_j-1}, \tag{3}$$

$$f(t; \boldsymbol{\theta}) = h(t; \boldsymbol{\theta}) \cdot R(t; \boldsymbol{\theta}), \tag{4}$$

where $\boldsymbol{\theta} = (k_1, \lambda_1, \ldots, k_m, \lambda_m)$ is the full parameter vector.

The shape parameter governs the failure mode: $k_j < 1$ indicates decreasing hazard (infant mortality), $k_j = 1$ gives the exponential (constant hazard), and $k_j > 1$ gives increasing hazard (wear-out). The mean time to failure is $\text{MTTF}_j = \lambda_j \Gamma(1 + 1/k_j)$.

2.2 Masked and Censored Data

We observe n independent system lifetimes under two forms of incomplete information:

- **Right-censoring**: Some systems are still operating at the end of the observation period. For system i, $\delta_i = 1$ indicates failure and $\delta_i = 0$ indicates censoring.
- **Masking**: For failed systems, only a candidate set $C_i \subseteq \{1, \ldots, m\}$ of possible failure causes is observed, not the true failed component K_i. The candidate set always contains the true cause ($K_i \in C_i$), and the masking mechanism is non-informative about $\boldsymbol{\theta}$ [2].

The log-likelihood for the complete sample is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n}\left[(1-\delta_i)\log R(t_i; \boldsymbol{\theta}) + \delta_i \log\left(\sum_{j \in C_i}\frac{h_j(t_i)}{h(t_i; \boldsymbol{\theta})} \cdot f(t_i; \boldsymbol{\theta})\right)\right], \tag{5}$$

which is maximized numerically using the L-BFGS-B algorithm [20] with analytical gradients. For details, see [2].

2.3 The Common-Shape Model

The *common-shape model* (reduced model) constrains all components to share a single shape parameter k, with parameter vector $\boldsymbol{\theta}_R = (k, \lambda_1, \ldots, \lambda_m)$. This reduces the parameter count from $2m$ to $m + 1$.

Property 2.1 (Weibull Closure [1]). *Let $T_1, \ldots, T_m$ be independent with $T_j \sim$ Weibull(k, λ_j). Then $T = \min\{T_1, \ldots, T_m\} \sim$ Weibull(k, λ_s) where $\lambda_s = (\sum_{j=1}^{m} \lambda_j^{-k})^{-1/k}$.*

Remark 2.1 (Uniqueness)**.** *The common-shape constraint is the* only *single-parameter restriction on the full model that yields a Weibull system lifetime. This follows from the linear independence of* $\{e^{\alpha u} : \alpha \in \mathbb{R}\}$*: if any* $k_i \neq k_j$*, the system cumulative hazard* $\sum_j (t/\lambda_j)^{k_j}$ *cannot equal* $(t/\lambda_s)^{k_s}$ *for any* (k_s, λ_s)*.*

This Weibull closure provides analytical tractability (closed-form MTTF, reliability function, hazard), interpretability (single failure mode), and computational efficiency from fewer parameters. However, the key practical question is not whether the property holds exactly, but whether predictions from the common-shape model are adequate when shapes are *approximately* equal.

2.4 Model Hierarchy

Several nested reductions of the full model are possible (Table 1):

Table 1: Hierarchy of Nested Models for Weibull Series Systems

Model	Parameters	Count	System Weibull?	Physical Meaning
Full	$(k_1, \lambda_1, \ldots, k_m, \lambda_m)$	$2m$	No	General
Common shape	$(k, \lambda_1, \ldots, \lambda_m)$	$m+1$	**Yes**	Same aging, different durability
Common scale	$(k_1, \ldots, k_m, \lambda)$	$m+1$	No	Different aging, same durability
Fully homogeneous	(k, λ)	2	Yes	Identical components
Exponential	$(\lambda_1, \ldots, \lambda_m)$, $k = 1$	m	Yes	No aging

The common-shape model is the natural "Goldilocks" reduction for three reasons. *Mathematically*: it is the unique single-parameter constraint preserving the Weibull property (Property 2.1). *Physically*: in well-designed systems, components are manufactured to similar quality standards (similar aging $\rightarrow$ similar k_j) but differ in size, load, and materials (different durability $\rightarrow$ different λ_j). The common-scale model would assert similar lifetimes but fundamentally different failure mechanisms, which rarely occurs. *Empirically*: the fully homogeneous model ($k_j = k$, $\lambda_j = \lambda$) is rejected at rates of 58–99% for our baseline system even though scale CV is only 5.4%, confirming that scale heterogeneity is real and important.

2.5 Divergence Metric

We measure departure from shape homogeneity using the coefficient of variation:

$$\mathrm{CV}_k = \frac{\mathrm{sd}(k_1, \ldots, k_m)}{\mathrm{mean}(k_1, \ldots, k_m)}. \tag{6}$$

A system with $\mathrm{CV}_k = 0$ has perfectly homogeneous shapes. Our baseline system (Table 2) has $\mathrm{CV}_k \approx 4\%$.

2.6 Baseline System and Simulation Design

Throughout this paper, we use a 5-component baseline system from [12]:

All simulations use masking probability $p = 0.215$ and censoring quantile $q = 0.825$ unless otherwise stated. To vary shape heterogeneity, we generate m shapes uniformly spaced about a mean of 1.18 with a specified target CV. Because the shapes are discrete, the realized CV differs slightly from the target (e.g., target CV = 10% yields actual CV $\approx$ 13.7%). All tables and figures report the actual CV; approximate values are used in prose for readability. All simulations use the

Table 2: Baseline 5-Component Series System

Component j	Shape k_j	Scale λ_j	MTTF$_j$
1	1.2576	994.37	≈ 913
2	1.1635	908.95	≈ 859
3	1.1308	840.11	≈ 799
4	1.1802	940.13	≈ 886
5	1.2034	923.16	≈ 866

`wei.series.md.c1.c2.c3` R package [21] with L-BFGS-B optimization [20] and analytical gradients. Appendix A examines how perturbations to individual component parameters affect MLE performance, and Appendix B confirms that estimation difficulty is inherent to series systems even without masking or censoring.

3 Consequence Analysis: Does Model Misspecification Matter?

The central question is not whether the common-shape model is statistically distinguishable from the full model, but whether it produces predictions that are wrong enough to affect engineering decisions. We quantify this by fitting the reduced model to data generated from heterogeneous-shape systems and measuring the resulting prediction errors.

3.1 Prediction Metrics

For each simulation replication, we compute three reliability metrics under both the fitted full model and the fitted reduced model, and compare to the ground truth:

1. **System MTTF**: $\text{MTTF} = \int_0^\infty R(t;\boldsymbol{\theta})\,dt$
2. **System reliability**: $R(t;\boldsymbol{\theta})$ evaluated at $t = \text{MTTF}/2$, MTTF, and $2 \cdot \text{MTTF}$
3. **Component failure probabilities**: $P_j = \Pr\{K_i = j\} = \int_0^\infty h_j(t)R(t)\,dt$

The relative bias of the reduced model is $(\hat{M}_R - M_{\text{true}})/M_{\text{true}}$ for each metric M.

3.2 Simulation Design

We varied shape CV across 9 levels (0–30%) and sample sizes $n \in \{100, 500, 1000, 5000\}$, with 500 replications per condition. At each condition, we fit both the full and reduced models and compute all prediction metrics. Non-convergent fits (typically 0–3% at $n = 100$ for moderate CV, increasing to approximately 10% at extreme CV levels (>40%); rates do not always decrease monotonically with n) are excluded from the analysis.

3.3 Results

Table 3 summarizes the reduced model's MTTF bias across shape CV and sample size. Figure 1 provides the corresponding visualization.

Three patterns emerge. First, the reduced model's MTTF bias is remarkably small: below 1% through shape CV $\approx 14\%$ at $n \geq 500$, and the 1% threshold is not crossed until CV $\approx 20\%$ (bias = 1.5%). Only at CV $> 27\%$ does the bias exceed 2.5%. Second, the bias is almost entirely

Table 3: Reduced Model MTTF Bias (%) and RMSE (in time units) by Actual Shape CV and Sample Size

	$n = 100$		$n = 500$		$n = 1000$		$n = 5000$	
CV (%)	Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
0.0	+0.4	20.4	−0.1	9.6	+0.0	6.8	−0.0	2.9
2.7	+0.3	21.0	−0.1	10.0	−0.1	6.9	+0.0	2.9
5.5	+0.8	21.2	+0.2	9.3	+0.1	6.6	+0.1	3.1
8.2	+0.7	20.5	+0.5	9.5	+0.1	6.6	+0.2	3.0
11.0	+1.1	21.5	+0.6	9.8	+0.2	6.7	+0.2	3.0
13.7	+0.8	21.8	+0.9	10.1	+0.5	7.0	+0.6	3.3
20.5	+1.7	23.9	+1.5	10.5	+1.3	7.7	+1.3	4.2
27.4	+3.2	23.9	+2.9	12.5	+2.5	9.2	+2.6	6.4
41.1	+10.5	37.4	+10.1	23.1	+9.8	20.8	+9.3	18.2

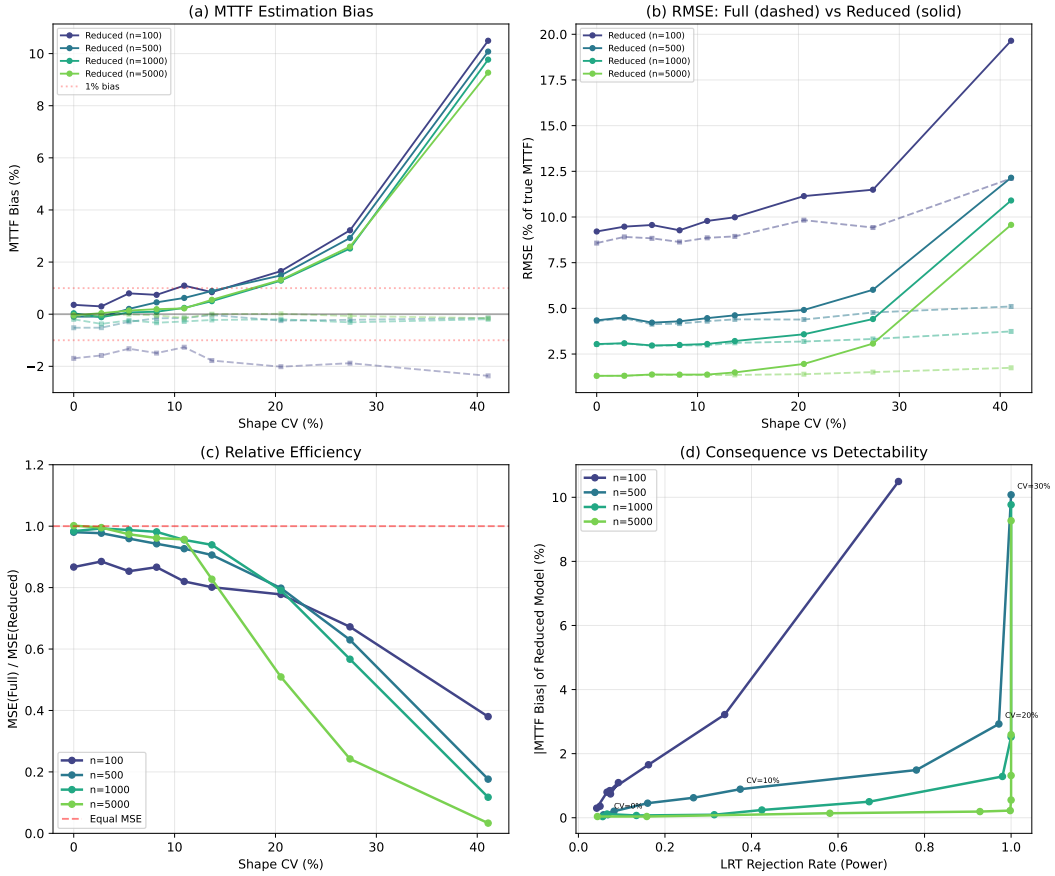

Figure 1: Consequence analysis. (a) MTTF bias: dashed lines show the full model (near zero at all CVs), solid lines show the reduced model (growing with CV). (b) RMSE comparison as percentage of true MTTF. (c) Relative efficiency: the full model has lower MSE at all CVs, increasingly so for higher CV. (d) Consequence vs detectability: the LRT rejects the reduced model only when bias becomes large.

positive—the reduced model systematically overestimates MTTF because the common-shape fit overestimates the weaker components. Third, the bias is largely independent of sample size: a 50-fold increase from $n = 100$ to $n = 5000$ changes the bias at CV $\approx 27\%$ from $+3.2\%$ to $+2.6\%$, confirming that this is a specification bias, not a finite-sample artifact.

The reliability function $R(t)$ evaluated at the system MTTF shows analogous behavior. The reduced model's $R(\text{MTTF})$ bias is below 0.5% for CV $\leq 10\%$ and grows to approximately -4% at CV $= 30\%$. The negative sign indicates that the reduced model underestimates reliability at the MTTF, consistent with the Weibull closure property forcing a compromise shape.

3.4 Discussion

A surprising finding is that the full model has lower MSE than the reduced model at practical sample sizes ($n \leq 1000$)—even when the reduced model is correctly specified (CV $= 0$). A bias-variance decomposition (Figure 2) reveals the mechanism: at $n = 100$ and CV $= 0$, the full model's MTTF variance (348) is substantially lower than the reduced model's (417), despite having more parameters. The traditional argument that parsimony reduces estimator variance holds for the shape parameters themselves, but the system MTTF is a nonlinear functional $\text{MTTF} = \int_0^\infty \exp\{-\sum_j (t/\lambda_j)^{k_j}\}\, dt$ that can amplify variance differently under constrained versus unconstrained parameterizations [22]. The constraint $k_1 = \cdots = k_m$ forces all shape estimation error into a single degree of freedom, which then propagates through the nonlinear MTTF integral with greater amplification than the distributed errors of the unconstrained estimator. At $n = 5000$ the two models achieve approximately equal MSE as the variance contribution diminishes.

This has a practical implication: *the traditional argument for the reduced model—lower variance from fewer parameters—does not hold for system-level predictions.* The case for the common-shape model rests instead on interpretability, the Weibull closure property, and the empirical observation that its predictions are accurate enough for engineering use when shape heterogeneity is modest.

Figure 1(d) reveals the relationship between detectability and consequence: the LRT rejection rate and the reduced model's bias both increase with CV, but they track different curves. At $n = 500$: at CV $\approx 14\%$, the LRT rejection rate is 37% while the MTTF bias is only 0.9%—the test has substantial probability of rejecting a model whose predictions are adequate. At CV $\approx 21\%$, the test rejects 97% of the time and the bias is 1.5%. At CV $\approx 27\%$, the bias reaches 2.9%, warranting the switch to the full model.

4 Likelihood Ratio Testing

The consequence analysis shows *when misspecification matters* for predictions. This section addresses a complementary question: *when can the data distinguish* between the full and reduced models? We employ the likelihood ratio test (LRT) with statistic

$$\Lambda = -2(\ell(\hat{\boldsymbol{\theta}}_R) - \ell(\hat{\boldsymbol{\theta}}_F)), \tag{7}$$

which is asymptotically χ^2_{m-1} under the null hypothesis $H_0 : k_1 = \cdots = k_m$ [16].

4.1 Type I Error Validation

Table 4 presents rejection rates at $\alpha = 0.05$ under perfect homogeneity ($\text{CV}_k = 0$). All rates are consistent with the nominal level, confirming the χ^2_{m-1} approximation. The χ^2 approximation is asymptotic; at $n = 100$ with $m - 1 = 4$ degrees of freedom, finite-sample deviations may occur, though our Type I error validation confirms adequate calibration across all sample sizes tested.

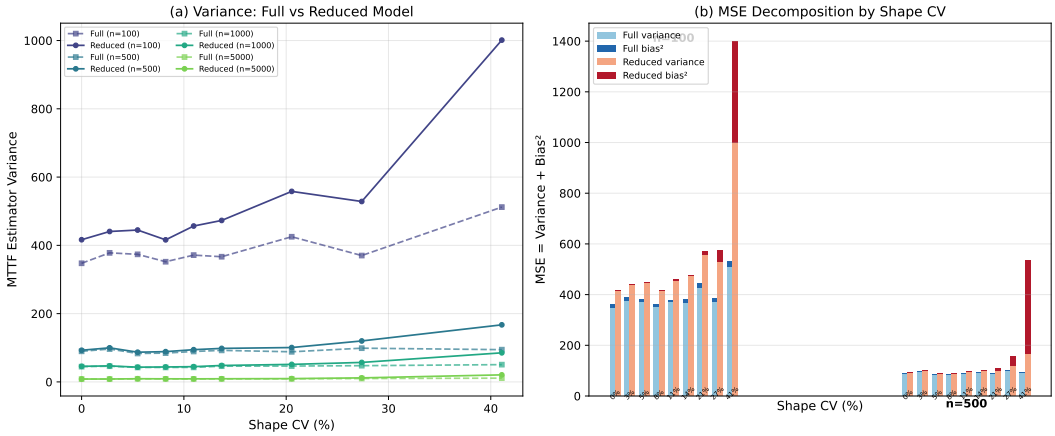

Figure 2: Bias-variance decomposition of MTTF MSE. (a) Variance comparison: the full model has lower MTTF variance than the reduced model at all CVs, despite having more parameters. (b) MSE decomposition: the reduced model's MSE is dominated by variance at low CV, then by bias^2 at high CV.

Table 4: LRT Type I Error Rate Under Perfect Homogeneity ($CV_k = 0$)

Sample Size n	Rejection Rate	95% CI	Status
100	0.054	[0.037, 0.077]	OK
500	0.056	[0.039, 0.080]	OK
1000	0.046	[0.031, 0.068]	OK
5000	0.068	[0.049, 0.094]	OK
10000	0.046	[0.031, 0.068]	OK

4.2 Power Analysis

Figure 3 and Table 5 present the LRT rejection rate (power) as a function of shape CV for different sample sizes.

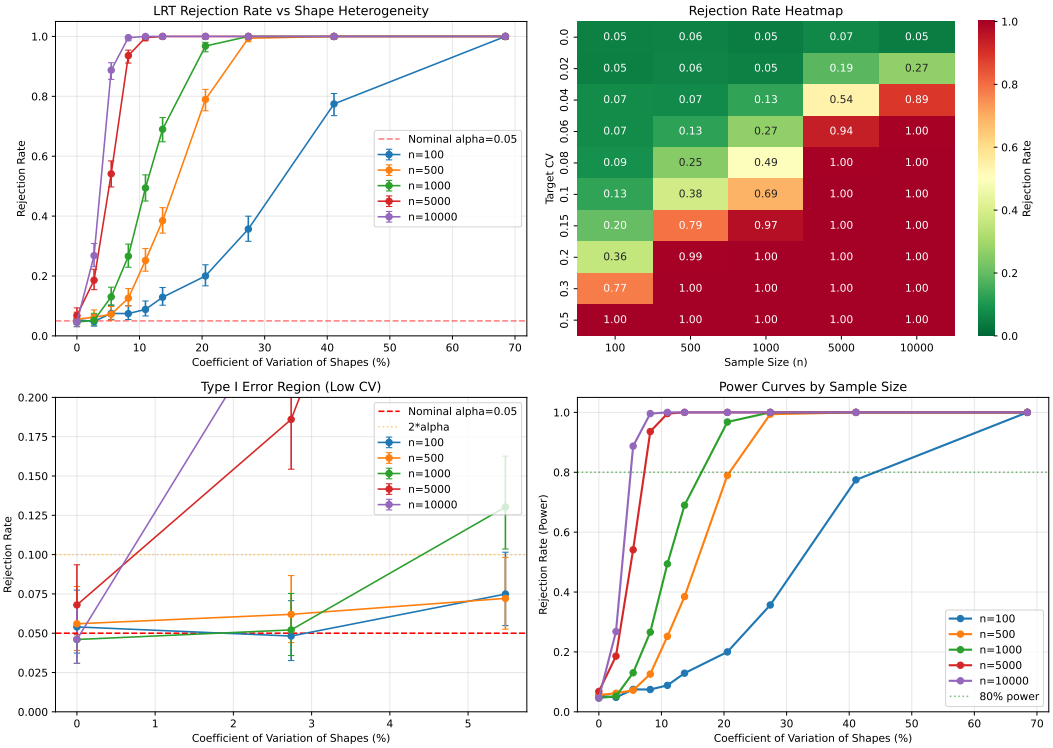

Figure 3: LRT power as a function of shape parameter divergence. Top left: Power curves by sample size. Top right: Heatmap of rejection rates. Bottom left: Type I error region (low CV). Bottom right: Power curves (clean view, no CIs).

Table 5: LRT Rejection Rate by Divergence Level and Sample Size

CV (%)	$n = 100$	$n = 500$	$n = 1000$	$n = 5000$	$n = 10000$
0.0	0.054	0.056	0.046	0.068	0.046
2.7	0.048	0.062	0.052	0.186	0.268
5.5	0.075	0.072	0.130	0.541	0.887
8.2	0.074	0.126	0.266	0.936	0.996
11.0	0.088	0.252	0.494	0.996	1.000

The key observation: at CV $\leq$ 5%, the LRT has almost no power at practical sample sizes ($n \leq 1000$). Even at $n = 10{,}000$, the rejection rate at CV = 2.7% is only 27%. Combined with the consequence analysis showing that prediction bias is negligible in this regime, the case for the common-shape model is strong.

4.3 Factors Affecting Power

We examined how masking probability, censoring level, and system complexity affect LRT performance. These simulations used the baseline well-designed system (CV $\approx$ 4%).

Masking probability. Figure 4 shows that higher masking reduces power. At $n = 5000$, the rejection rate drops from 48% at $p = 0.05$ to 8% at $p = 0.70$. Masking reduces the information about which component caused each failure.

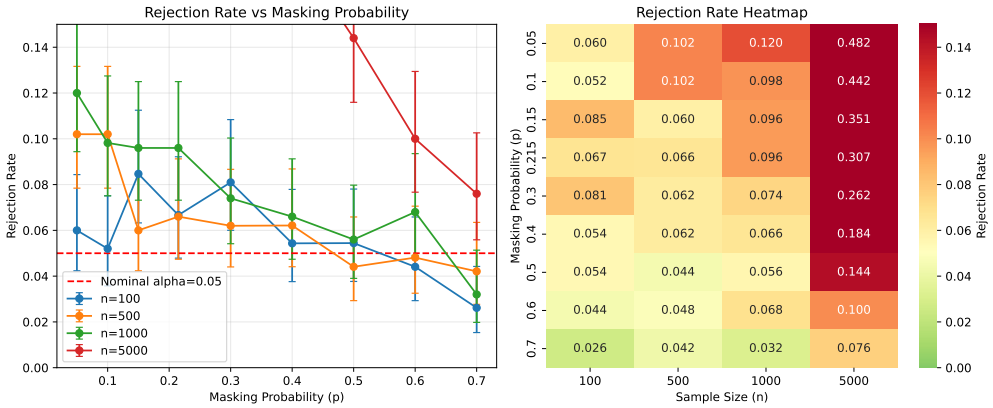

Figure 4: Effect of masking probability on LRT rejection rate.

Censoring level. Figure 5 shows that heavier censoring (lower q) reduces power. At $n = 5000$, the rejection rate increases from 17% at $q = 0.5$ to 43% at $q = 1.0$. Censored observations provide reliability information but no failure attribution.

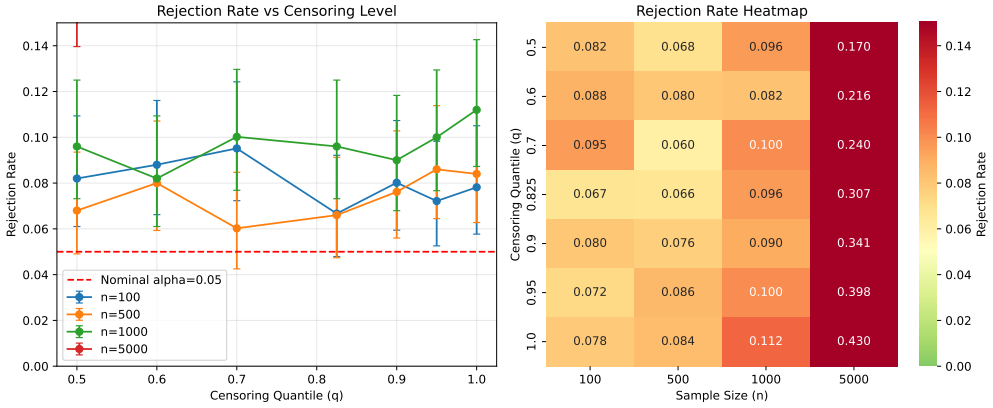

Figure 5: Effect of censoring quantile on LRT rejection rate.

System complexity. Figure 6 shows that power decreases with the number of components. At $n = 5000$, the rejection rate drops from 77% at $m = 2$ to 16% at $m = 8$. The LRT has $m-1$ degrees of freedom, so larger systems have higher critical values, and failure information is distributed across more parameters. Note that the shape CV decreases slightly from 5.5% ($m = 2$) to 3.5% ($m = 8$) in our design, which partially confounds the effect of system complexity with reduced heterogeneity.

Of the data quality factors, masking has the larger impact: increasing masking probability from 0.05 to 0.70 reduces power by a factor of 6, while removing censoring entirely increases power by a factor of 2.5. Importantly, neither factor inflates the Type I error rate—the test remains well-calibrated throughout.

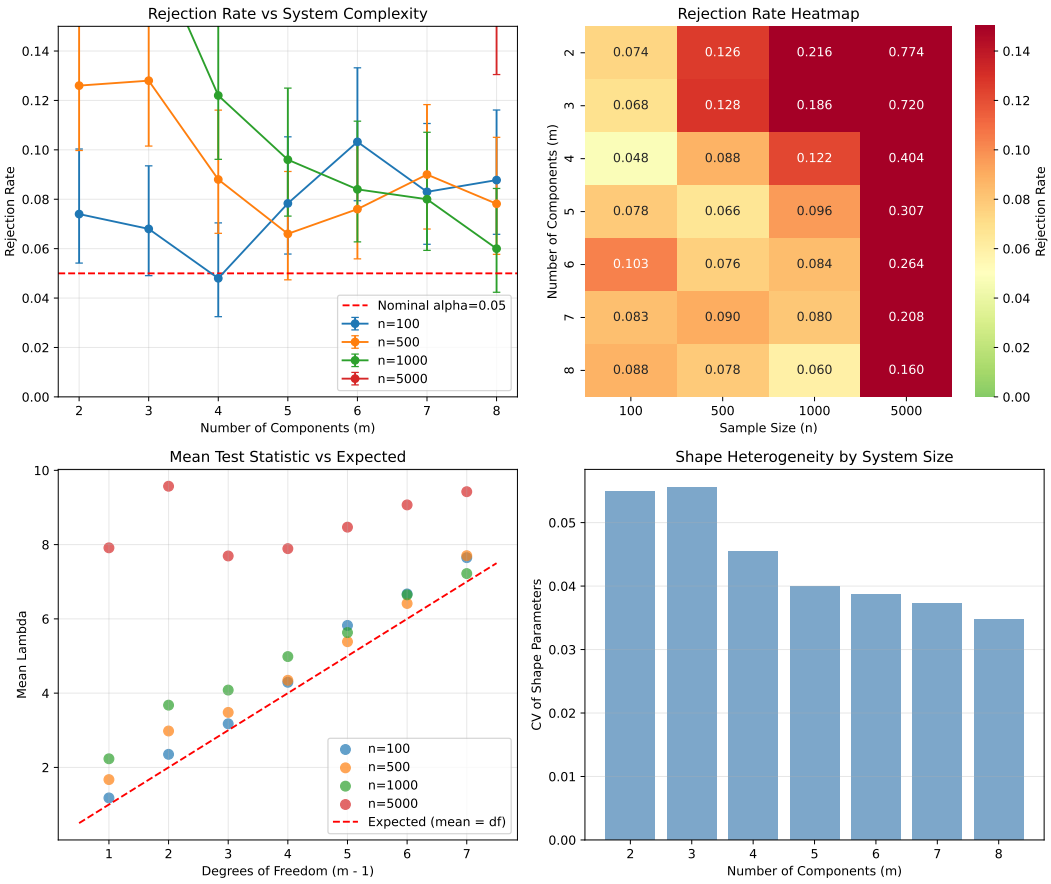

Figure 6: Effect of number of components on LRT rejection rate.

4.4 Comparison with Information Criteria

We compare the LRT with AIC [17] and BIC [18]. For two competing models with log-likelihoods ℓ_F ($2m$ parameters) and ℓ_R ($m+1$ parameters):

$$\text{AIC} = -2\ell + 2k, \tag{8}$$

$$\text{BIC} = -2\ell + k \ln n, \tag{9}$$

where k is the number of model parameters ($2m$ for the full model, $m+1$ for the reduced model).

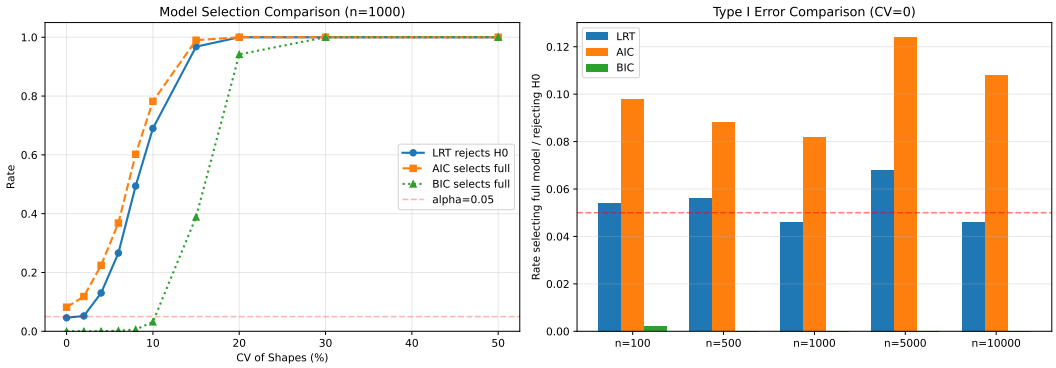

Figure 7: Model selection rates for LRT ($\alpha = 0.05$), AIC, and BIC vs shape CV.

Figure 7 and Table 6 summarize the comparison. Note that AIC and BIC are model selection criteria rather than hypothesis tests; the "Type I error" rates reported here represent the proportion of times each criterion selects the full model under perfect homogeneity, enabling a calibration comparison but not a direct power comparison. The LRT is well-calibrated (4.6–6.8%), AIC is liberal (8.2–12.4%, roughly 2× the nominal rate), and BIC is over-conservative (0–0.2%). For formal hypothesis testing of shape homogeneity, the LRT provides the best-calibrated Type I error while maintaining good power. AIC may be useful for exploratory analysis where false positives are less costly than missed heterogeneity. BIC's conservatism makes it unsuitable for detecting the subtle heterogeneity characteristic of well-designed systems.

Table 6: Type I Error Comparison Under Perfect Homogeneity ($\text{CV}_k = 0$)

Criterion	Selection Rate Range	Behavior
LRT ($\alpha = 0.05$)	4.6–6.8%	Well-calibrated
AIC	8.2–12.4%	Liberal ($\approx 2\times$ nominal)
BIC	0–0.2%	Over-conservative

5 Adaptive Model Selection

The preceding sections establish *when* the common-shape model is adequate (consequence analysis) and *whether* the data can distinguish it from the full model (LRT). This section addresses the practical question: given data from an unknown system, how should a practitioner choose between models?

5.1 Procedure

We evaluate a natural LRT-based adaptive procedure:

1. Fit the full model ($2m$ parameters) to obtain $\hat{\boldsymbol{\theta}}_F$.
2. Fit the reduced model ($m+1$ parameters) to obtain $\hat{\boldsymbol{\theta}}_R$.
3. Compute the LRT statistic $\Lambda = -2(\ell(\hat{\boldsymbol{\theta}}_R) - \ell(\hat{\boldsymbol{\theta}}_F))$.
4. If $\Lambda < \chi^2_{m-1,1-\alpha}$ (i.e., $p > \alpha$), use the reduced model; otherwise use the full model.

This procedure inherits the well-calibrated Type I error of the LRT (Table 4): under the null, the reduced model is selected with probability $1-\alpha$.

We also compare with a simpler CV-based approach: fit the full model, compute $\widehat{\text{CV}}_k = \text{sd}(\hat{k}_1, \ldots, \hat{k}_m)/\text{mean}(\hat{k}_1, \ldots, \hat{k}_m)$, and select the reduced model if $\widehat{\text{CV}}_k < \tau$ for a threshold τ.

5.2 Simulation Design

We compare four strategies:

1. **Always-full**: use full model estimates regardless
2. **Always-reduced**: use reduced model estimates regardless
3. **Adaptive (LRT)**: select model via the LRT at $\alpha = 0.05$
4. **Adaptive (CV)**: select model via estimated CV threshold

We evaluate on MTTF estimation accuracy (MSE, bias) across shape CV from 0 to 20% and sample sizes $n \in \{100, 500, 1000\}$, with 500 replications per condition. Non-convergent fits are excluded (typically $< 2\%$ of replications).

5.3 Results

Table 7 presents the RMSE of each strategy as a percentage of the true system MTTF, along with the fraction of replications where the LRT-based adaptive procedure selects the reduced model. Figure 8 provides the corresponding visualization.

Table 7: Adaptive Model Selection: RMSE (% of True MTTF) and Selection Rate

CV (%)	n	Full	Reduced	Adaptive	Sel. Red. (%)
0.0	100	9.3	10.1	10.0	93
	500	4.6	4.7	4.7	95
	1000	3.0	3.0	3.0	94
5.5	100	9.3	10.1	10.1	95
	500	4.1	4.2	4.2	92
	1000	3.0	3.0	3.1	89
13.7	100	9.5	10.4	10.3	89
	500	4.3	4.5	4.4	60
	1000	3.0	3.2	3.0	30
20.5	100	9.5	10.7	10.4	83
	500	4.4	4.9	4.5	21
	1000	3.2	3.6	3.2	1
27.4	100	9.9	12.8	11.6	65
	500	4.6	5.6	4.7	1
	1000	3.3	4.4	3.3	0

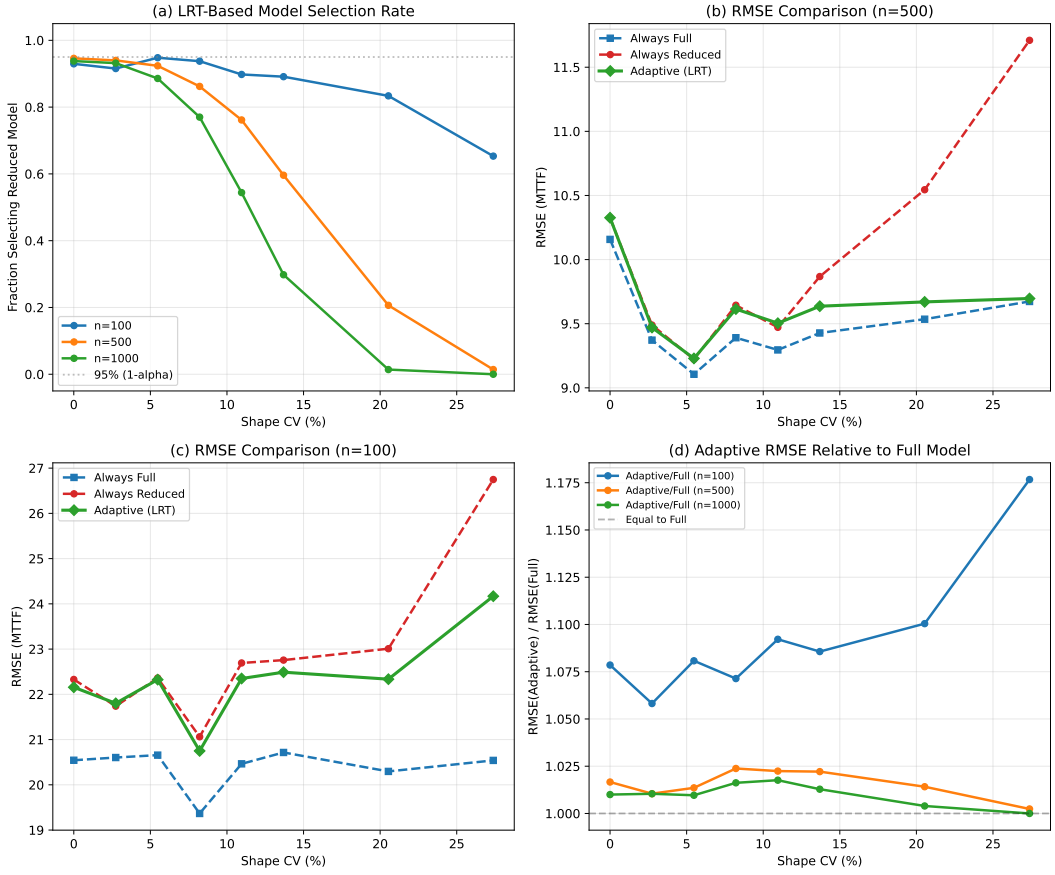

Figure 8: Adaptive model selection. (a) Selection rate: the LRT correctly selects the reduced model near 95% at CV = 0 and transitions smoothly to the full model as CV increases, with the transition point shifting left with sample size. (b)–(c) RMSE comparison at $n = 500$ and $n = 100$: the adaptive procedure (green) tracks the better of the two static strategies across all CVs. (d) Relative RMSE: the adaptive procedure's overhead ranges from 8–18% at $n = 100$ and is under 2.5% at $n = 500$.

The adaptive procedure exhibits three regimes. At **low CV** ($\leq 5\%$), the procedure selects the reduced model in $> 90\%$ of replications across all sample sizes, achieving RMSE comparable to the always-reduced strategy. At **moderate CV** (5–14%), the selection rate depends strongly on sample size: at $n = 100$ the procedure still selects reduced $\sim 90\%$ of the time (because the LRT lacks power), while at $n = 1000$ the selection rate drops to 30% at CV $\approx 14\%$. At **high CV** ($> 20\%$), the procedure overwhelmingly selects the full model at $n \geq 500$, protecting against the large bias of the reduced model.

The RMSE overhead of the adaptive procedure relative to the always-full strategy is at most 2.4% at $n = 500$ and converges to zero at large n and high CV. At $n = 100$, the overhead ranges from 8% to 18%, with the largest values occurring at high CV where the reduced model has substantial MTTF bias. However, even this "failure" is partially benign: the consequence analysis (Table 3) shows that the reduced model's bias remains moderate at the CVs where the LRT lacks power to reject it.

5.4 Discussion

The adaptive procedure succeeds because of a fortuitous alignment: the LRT has low power precisely where the consequence of misspecification is small, and high power where the consequence is large. At $n = 500$ and CV $\approx 14\%$, the adaptive procedure selects the reduced model 60% of the time; the bias of the reduced model at this CV is only 0.9%, so these "incorrect" selections carry negligible penalty. At CV $\approx 21\%$, the procedure selects the full model 79% of the time, and at CV $\approx 27\%$ it selects the full model 99% of the time, avoiding the 2.9% bias.

The CV-based adaptive procedure performs similarly to the LRT-based approach but requires choosing a threshold τ that does not automatically adapt to sample size or data quality. The LRT is preferred because its critical value incorporates the degrees of freedom and sample size naturally.

From a practical standpoint, the adaptive procedure adds no cost beyond fitting both models. Since both fits are required to compute the LRT statistic, the procedure is a free byproduct of the testing framework described in Section 4.

6 Conclusion

This paper addressed a fundamental question in reliability engineering: when can practitioners safely simplify a Weibull series system model by assuming common component shapes? The answer is structured around a quantitative alignment between prediction consequence and statistical detectability.

Bias and power are quantitatively aligned. The common-shape model's MTTF bias grows super-linearly with shape heterogeneity—remaining below 1% through CV $\approx 15\%$—while the LRT's power grows super-linearly. The practical consequence: the test lacks power precisely where misspecification is harmless for engineering decisions, and has high power precisely where the bias warrants switching to the full model. An adaptive LRT-based procedure exploits this alignment, achieving RMSE within 2.5% of the always-full strategy at $n \geq 500$ while selecting the simpler model over 90% of the time when the data support it.

The traditional argument for the reduced model does not hold. Surprisingly, a bias-variance decomposition (Figure 2) reveals that the full model has lower MTTF variance than the reduced model even when the reduced model is correctly specified (CV = 0). The constraint

$k_1 = \cdots = k_m$ forces all shape estimation error into a single degree of freedom, amplifying its propagation through the nonlinear MTTF integral. The case for the common-shape model rests instead on interpretability, the Weibull closure property, and the empirical demonstration that its predictions are engineering-adequate across a broad range of shape heterogeneity.

The LRT is well-calibrated; information criteria are not. Among model selection criteria, the LRT provides well-calibrated Type I error (4.6–6.8% at nominal $\alpha = 0.05$), while AIC is liberal ($\approx$2$\times$ nominal false positive rate) and BIC is over-conservative ($\leq$0.2%). The LRT has negligible power at CV below 5% for $n \leq 1{,}000$; even at $n = 10{,}000$, the rejection rate at CV $\approx 3\%$ is only 27%.

Data quality effects. Higher masking probabilities and heavier censoring reduce the power of all model selection methods by diminishing the information available for discriminating between models, but they do not inflate the Type I error rate. Of the two, masking has the larger impact (6$\times$ reduction in power from $p = 0.05$ to $p = 0.70$, vs 2.5$\times$ from $q = 0.50$ to $q = 1.00$).

Limitations. Our simulations use uniformly spaced shapes about a common mean. Real systems may exhibit asymmetric or clustered heterogeneity patterns. The single baseline system (5 components, moderate masking and censoring) may not capture all configurations of practical interest.

Future directions. Three extensions merit investigation. First, the common-shape assumption could be tested for non-Weibull distributions where analogous closure properties may or may not hold. Second, optimal experimental design for model selection would provide guidance on sample size and diagnostic effort. Third, extending to parallel and k-out-of-n configurations would broaden applicability.

A MLE Sensitivity to System Design Perturbations

This appendix examines how deviations from the baseline system configuration affect MLE performance, motivating the model selection analysis in the main text.

A.1 Effect of Scale Parameter Perturbation

Figure 9 shows the effect of varying component 3's scale parameter λ_3 (mapped to MTTF_3) on MLE dispersion, bias, and coverage probability.

As component 3's MTTF decreases below other components, it causes more system failures, providing more information for its own parameter estimation (lower dispersion, better coverage) at the expense of other components.

A.2 Effect of Shape Parameter Perturbation

Figure 10 shows the effect of varying component 3's shape parameter k_3 on MLE performance. The relationship between k_3 and the probability of component 3 causing system failure is complex: components with $k < 1$ exhibit high early hazard and dominate system failures despite having higher MTTFs.

These results confirm that (1) shape parameters are inherently harder to estimate than scale parameters, and (2) estimator precision depends strongly on component failure probability. Both

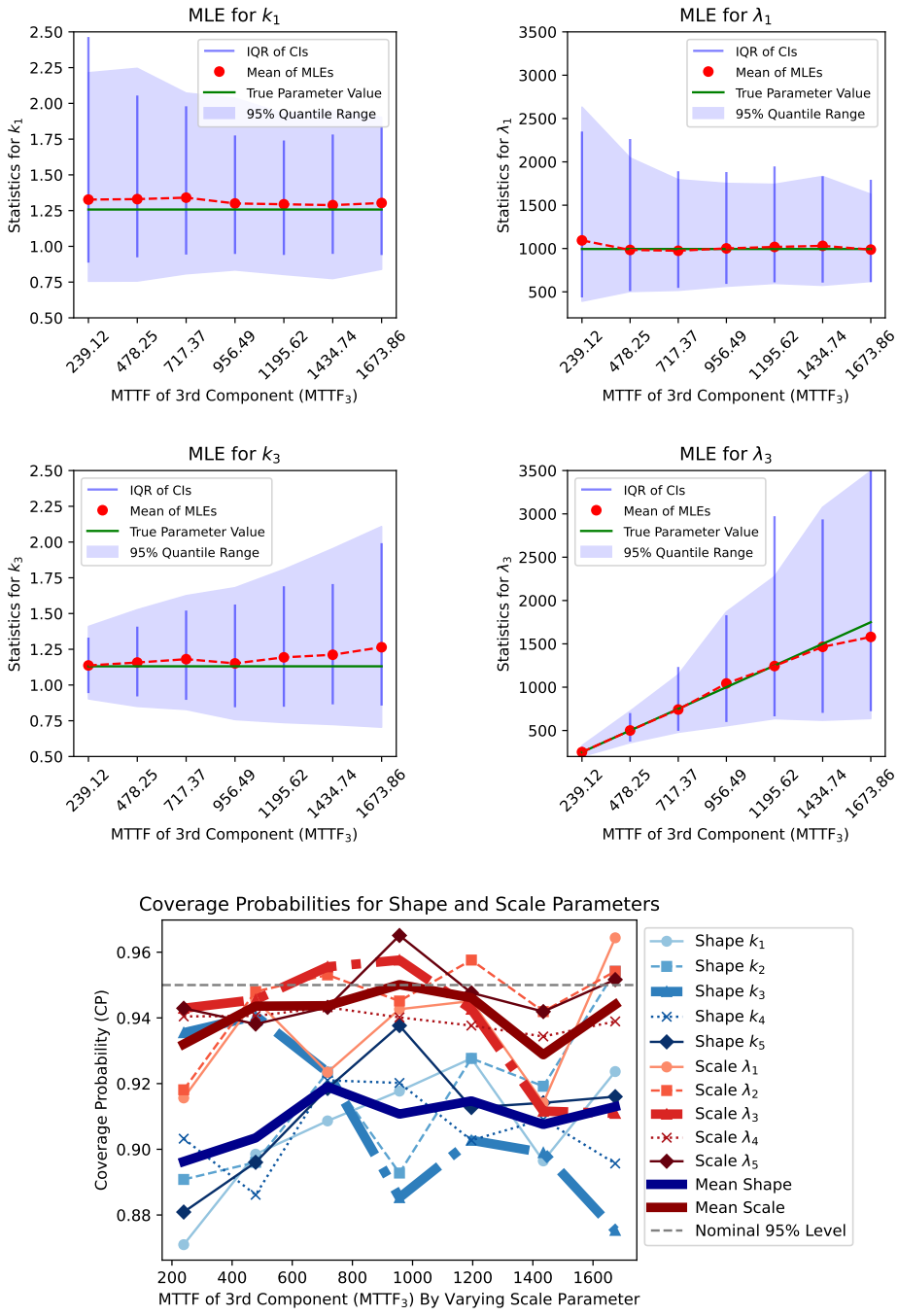

Figure 9: MLE performance vs MTTF of component 3 (by varying scale λ_3). Components with smaller MTTF cause more system failures and are estimated more precisely.

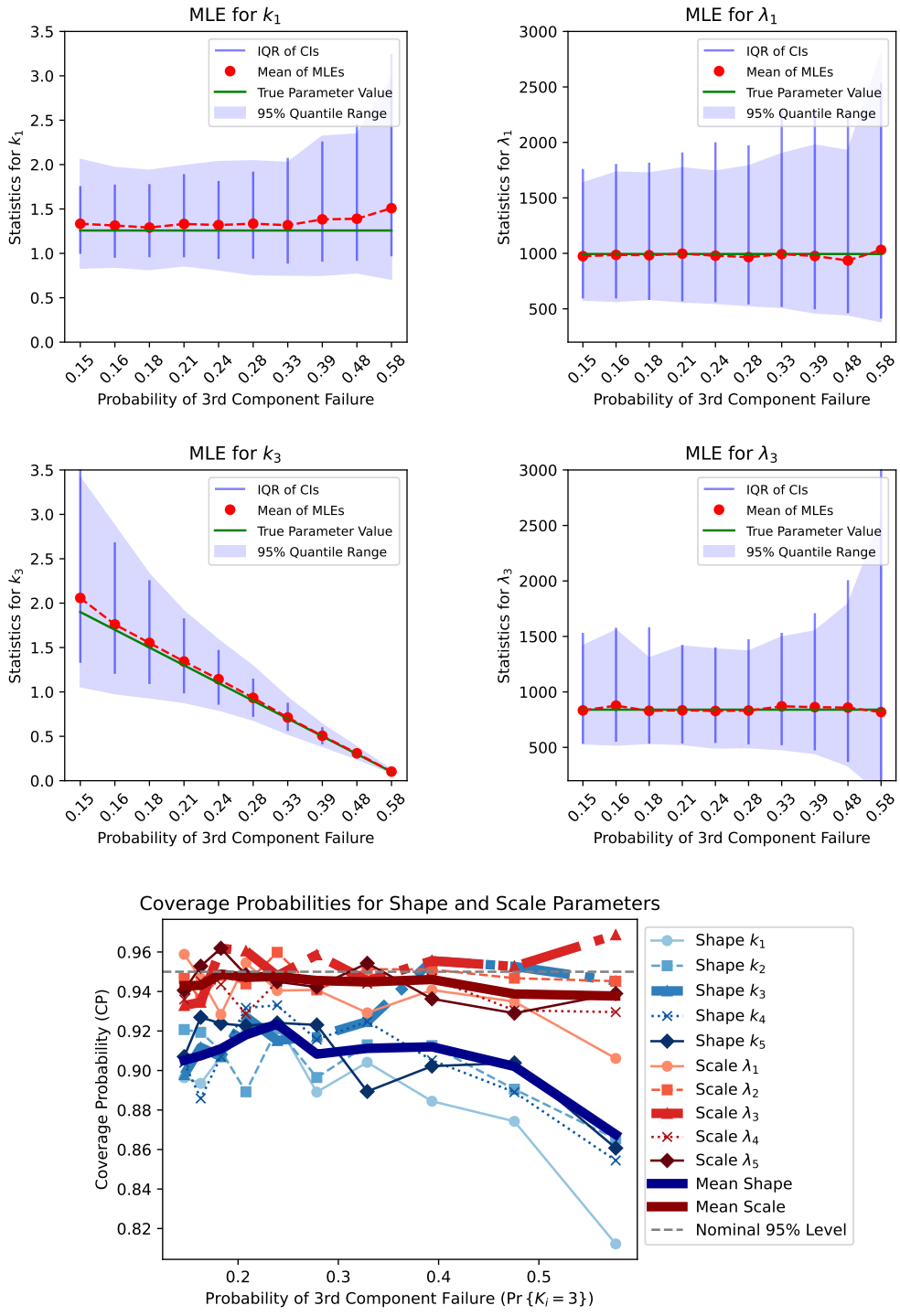

Figure 10: MLE performance vs probability of component 3 failure (by varying shape k_3). Shape parameters are harder to estimate than scale parameters, particularly when failure probability is low.

observations motivate the common-shape model: by pooling shape information across components, the reduced model effectively increases the sample size for the shared shape parameter.

B Ideal Case Analysis

Figure 11 examines MLE behavior under ideal conditions—no masking ($p = 0$) and no censoring ($q = 1$)—using a 2-component system with $n = 100$. Even with perfect information, shape parameter estimates exhibit bias for components with low failure probabilities, confirming that estimation difficulty is inherent to series systems, not solely a consequence of masked or censored data.

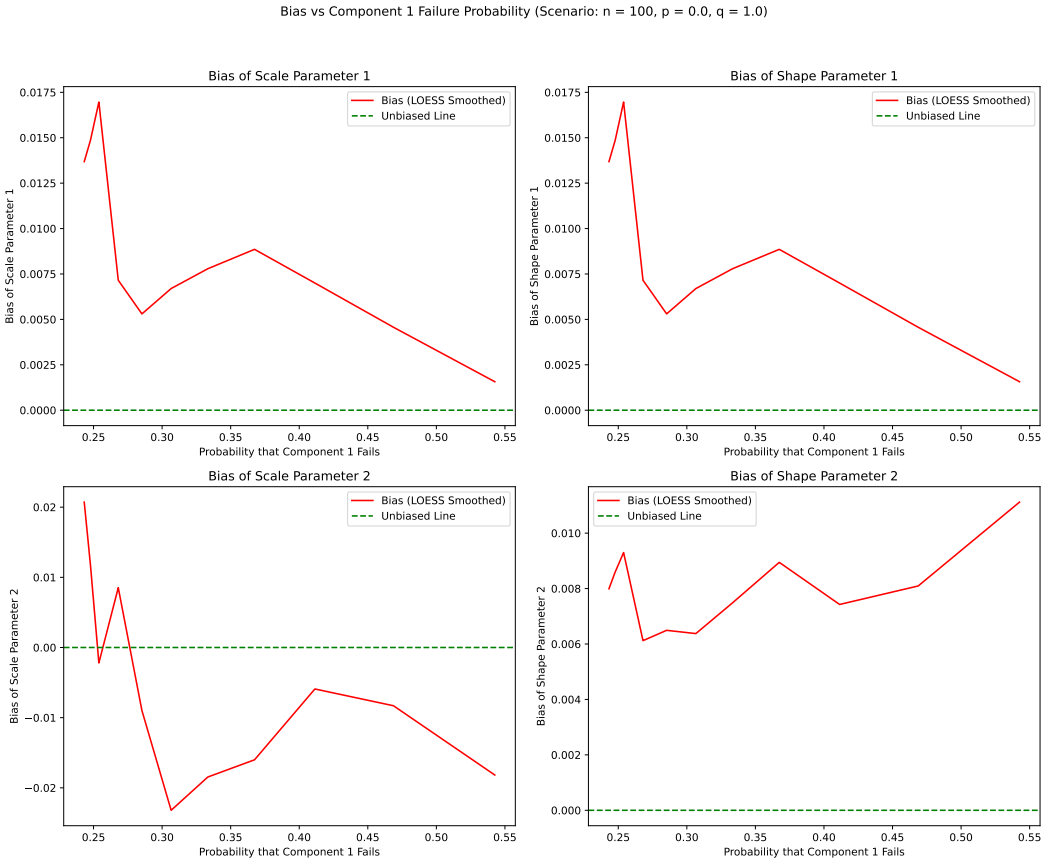

Figure 11: MLE behavior under ideal conditions (no masking, no censoring) for a 2-component series system with $n = 100$.

References

[1] R. E. Barlow and F. Proschan, *Statistical Theory of Reliability and Life Testing: Probability Models*. New York: Holt, Rinehart and Winston, 1975.

[2] A. Towell, "Reliability estimation in series systems: Maximum likelihood techniques for right-censored and masked failure data," 2023, [Online; accessed 2026-02-18]. [Online]. Available: https://github.com/queelius/reliability-estimation-in-series-systems

[3] J. Usher and T. Hodgson, "Maximum likelihood analysis of component reliability using masked system life-test data," *IEEE Transactions on Reliability*, vol. 37, no. 5, pp. 550–555, 1988.

[4] D. Lin, J. Usher, and F. Guess, "Exact maximum likelihood estimation using masked system data," *IEEE Transactions on Reliability*, vol. 42, no. 4, pp. 631–635, 1993.

[5] ——, "Bayes estimation of component-reliability from masked system-life data," *IEEE Transactions on Reliability*, vol. 45, no. 2, pp. 233–237, jun 1996.

[6] J. Usher, "Weibull component reliability-prediction in the presence of masked data," *IEEE Transactions on Reliability*, vol. 45, no. 2, pp. 229–232, jun 1996. [Online]. Available: http://doi.org/10.1109/24.510806

[7] F. M. Guess and J. S. Usher, "An iterative approach for estimating component reliability from masked system life data," *Quality and Reliability Engineering International*, vol. 5, no. 4, pp. 257–261, oct 1989.

[8] A. M. Sarhan, "Reliability estimations of components from masked system life data," *Reliability Engineering & System Safety*, vol. 74, no. 1, pp. 107–113, Oct. 2001.

[9] ——, "Parameter estimations in linear failure rate model using masked data," *Applied Mathematics and Computation*, vol. 151, no. 1, pp. 233–249, mar 2004.

[10] Z. Tan, "Estimation of component failure probability from masked binomial system testing data," *Reliability Engineering & System Safety*, vol. 88, no. 3, pp. 301–309, jun 2005.

[11] ——, "Estimation of exponential component reliability from uncertain life data in series and parallel systems," *Reliability Engineering & System Safety*, vol. 92, no. 2, pp. 223–230, feb 2007.

[12] H. Guo, P. Niu, and F. Szidarovszky, "Estimating component reliabilities from incomplete system failure data," *Annual Reliability and Maintainability Symposium (RAMS)*, pp. 1–6, jan 2013.

[13] J. F. Lawless, *Statistical Models and Methods for Lifetime Data*, 2nd ed. Hoboken, NJ: John Wiley & Sons, 2003.

[14] W. Q. Meeker and L. A. Escobar, *Statistical Methods for Reliability Data*. New York: John Wiley & Sons, 1998.

[15] R. Craiu and T. Lee, "Model selection for the competing-risks model with and without masking," *Technometrics*, vol. 47, no. 4, pp. 457–467, 2005.

[16] S. S. Wilks, "The large-sample distribution of the likelihood ratio for testing composite hypotheses," *The Annals of Mathematical Statistics*, vol. 9, no. 1, pp. 60–62, 1938.

[17] H. Akaike, "A new look at the statistical model identification," *IEEE Transactions on Automatic Control*, vol. 19, no. 6, pp. 716–723, 1974.

[18] G. Schwarz, "Estimating the dimension of a model," *The Annals of Statistics*, vol. 6, no. 2, pp. 461–464, 1978.

[19] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd ed. New York: Springer, 2002.

[20] R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu, "A limited memory algorithm for bound constrained optimization," *SIAM Journal on Scientific Computing*, vol. 16, no. 5, pp. 1190–1208, 1995.

[21] A. Towell, *wei.series.md.c1.c2.c3: Estimating Reliability of Weibull Components in Series from Masked Data*, 2023, r package version 0.9.0. [Online; accessed 2026-02-18]. Available: https://github.com/queelius/wei.series.md.c1.c2.c3/. [Online]. Available: https://queelius.github.io/wei.series.md.c1.c2.c3/

[22] H. White, "Maximum likelihood estimation of misspecified models," *Econometrica*, vol. 50, no. 1, pp. 1–25, 1982.