10

Combining Model and Test Data for Optimal Determination of Percentiles and Allowables: CVaR Regression Approach, Part II

Stan Uryasev[1] and A. Alexandre Trindade[2]

[1] American Optimal Decisions, Inc.
and Department of Industrial and Systems Engineering
University of Florida
uryasev@ufl.edu
[2] Department of Statistics
University of Florida

Summary. This report makes a more detailed assessment of the CVaR regression method proposed in [3] for determining A-Basis and B-Basis allowables, and quantifying the impact of test data and different analytical models on failure load predictions. Although the method can in principle be applied to any desired quantile, we will focus only on the 10th (B-Basis) in this study. We consider failure data arising from two sources: (1) a controlled environment where data is simulated from different Weibull distributions; (2) a supplied dataset similar to that of [3] augmented with failure load predictions from two additional analytical models (Model S2 and Model S3). Using absolute deviation between true and estimated 10th percentiles and the CVaR regression goodness-of-fit measure introduced in [3] as accuracy-assessment criteria, the key findings are as follows. The accuracy of CVaR regression is relatively insensitive to the number of batches present, but fairly sensitive to the number of test points per batch[3]. There are diminishing benefits in using more than 10 batches, or more than 10 test points per batch, in any one application of CVaR regression. The estimates of A-basis and B-basis are fairly robust, in the sense that they are not severely affected by miscalibrations (biases or errors) in the analytical models. Among the analytical models used as the sole input with no (input) test data, the best performer is Model S, followed by Model S2. Model S3 is the worst performer. The models contribute substantially to percentile prediction when up to 3 test points are used as input. When 4 test points are used as input, the 3 models can be roughly equated to the input information provided by one additional test point.

Executive Summary

In [3], we proposed a coherent methodology for integrating various sources of variability on properties of materials in order to accurately predict percentiles of their failure load distribution. The approach, **CVaR regression**, involved the linear combination of factors that are associated with failure load, into a statistical regression or *factor model*. The method can be used for determining A-Basis and B-Basis allowables, and quantifying the impact of experimental (or test) data and different analytical models on failure load predictions. The present report builds on this work by considering failure data arising from two sources: (1) a controlled environment where data is simulated from different Weibull distributions with

[3] A "batch" is defined as any one single source of variability affecting the test point failure data.

parameters in ranges plausibly mimicking failure data from a supplied dataset similar to that of [3]; (2) a supplied dataset with failure load predictions from three analytical models.

Specific Objectives

1. Pooling of various sources of information of possibly different origins: models, experiments and expert opinions.
2. Develop a methodology for estimating percentiles of failure distributions and allowables.
3. Minimize amount of data needed for certification process.
4. Take into account various sources of uncertainty.
5. Validate the approach in a controlled statistical environment.
6. Demonstrate the efficacy of the approach with case studies.

Summary of Accomplished Tasks and Findings

1. Developed factor model for direct estimation of percentiles using:
   a) various sources of information (models, experiments, expert opinions); it is possible to quantify value of different sources of information.
   b) statistical characteristics: mean, st.dev., deviation CVaR, etc.
2. Simple, clear, computationally effective methodology enables the pooling of data across:
   a) many individual materials: relatively small requirements on size of datasets.
   b) various experimental setups: crediting simple experiments to more sophisticated (expensive) ones.
3. Developed CVaR statistical techniques for optimal estimation (weighting) of coefficients in the factor model, and corresponding confidence intervals for unknown parameters (A-basis and B-basis).
   a) Technique is new; developed in the framework of the AIM-C project.
   b) Approach was especially designed for estimating percentiles and constructing confidence intervals (A and B-basis).
   c) No distributional assumptions (such as normality) are made; method is nonparametric.
   d) CVaR deviation goodness-of-fit measure has exceptional mathematical and computational properties, allowing easy and efficient implementation of the methodology via linear programming: high speed of calculations; analysis of large datasets feasible; stable results.
4. Case studies were performed with simulated data.
   a) Minimal number of batches needed to calibrate the CVaR regression model was determined; sensitivity to this number.
   b) Minimal number of experimental test points per batch was determined; sensitivity to this number.
   c) Approximate relationship established between CVaR deviation error (observed) and true error (unobserved) in percentile estimation.
   d) Sensitivity of the approach to errors in analytical model information was assessed; methodology is robust to biases.
   e) 10th percentile estimates based on Model S individually, and on Model S plus 5 test points, are close to true values. B-basis values are also close to nominal values based on actual experiments.
5. Two case studies carried out for open-hole coupon dataset: estimation of 10th percentile and B-basis (failure data plus 3 analytical models).
   a) CVaR regression calculations provided plausible estimates of percentiles of failure load distribution.
   b) CVaR regression with analytical models only as predictor variables, provide plausible percentile and B-basis estimates, even in the absence of any experimental test data (used as predictors).
   c) Benefits of combining models for predicting percentiles were quantified.
   d) Benefits of combining models and experimental data were evaluated.
6. Our investigations provide compelling evidence that the methodology can effectively integrate modeling and experimental data, and reduce overall testing cost.

10.1 Introduction

In [3], we proposed a coherent methodology for integrating various sources of variability on properties of materials in order to accurately predict percentiles of their failure load distribution. The approach, **CVaR regression**, involved the linear combination of factors that are associated with failure load, such as might arise from analytical model and experimental test data, into a statistical regression or factor model. The method can be used for determining A-Basis and B-Basis allowables, and quantifying the impact of experimental test data and different analytical models on failure load characteristics. The purpose of the present report is to perform a more detailed evaluation of this CVaR regression method. The term **batch** is used here in the broad sense to denote any one single source of variability affecting the experimental test data. For example: stacking sequence of laminates, type of test, batch number, etc.

Using the terminology of [3], let Y_{ij} denote the j-th failure load value for the i-th batch, obtained from test data, $i = 1, \ldots, I$, $j = 1, \ldots, N_i$. Let (m_i, s_i) and (μ_i, σ_i) denote the sample mean and standard deviation for the test and model data, respectively, in batch i. There are several ways to form these summary statistics; two of these are outlined in Appendix A. Armed with these summary statistics, we can fit the regression or factor model:

$$Y_{ij} = c_0 + c_1\mu_i + c_2\sigma_i + c_3 m_i + c_4 s_i + \varepsilon_{ij}, \tag{10.1}$$

where the unknown regression coefficients $c_0, \ldots, c_4$ are to be estimated from the data by minimizing the deviation measure

$$\mathcal{P}_\tau^2 = \begin{cases} \mathrm{CVaR}_\tau^{\Delta}(\varepsilon), & \text{if } \tau \geq 0.5, \\ \mathrm{CVaR}_{1-\tau}^{\Delta}(-\varepsilon), & \text{if } \tau < 0.5. \end{cases} \tag{10.2}$$

Additional terms can be added to (10.1) in order to accommodate more factors, e.g. different analytical models. When the failure data is skewed, a better fit may be attained by symmetrizing it, e.g. taking logarithms. This will make the mean and variance more representative measures of location and dispersion within batches. More robust summary statistic measures, such as quantiles and tail means (CVaR), could also be used.

$\mathcal{P}_\tau^2$ can be used as an absolute measure of goodness-of-fit in CVaR regression, much like the Mean Square Error (MSE) of ordinary regression. For comparing two fitted models, M_2 to M_1 say, the quantity

$$\mathcal{P}_\tau^2(M_2, M_1) = \left[1 - \frac{\mathcal{P}_\tau^2(M_2)}{\mathcal{P}_\tau^2(M_1)}\right] 100\%,$$

measures the percentage improvement in fit obtained by using M_2 over M_1 (negative values meaning that the fit has worsened). This is exactly analogous to the *partial* R^2 of ordinary regression, which measures the percentage change in MSE(M_2) over MSE(M_1). If M_0 denotes the model with just an intercept term, then $\mathcal{P}_\tau^2(M_2, M_0)$ is exactly equivalent to the R^2 for model M_2.

A direct result of the CVaR regression fitting process is the following estimated equation for the τth quantile of the failure load, as a function of factors $\{\mu, \sigma, m, s\}$:

$$\hat{Q}_Y(\tau) = \hat{c}_0(\tau) + \hat{c}_1(\tau)\mu + \hat{c}_2(\tau)\sigma + \hat{c}_3(\tau)m + \hat{c}_4(\tau)s. \tag{10.3}$$

The calculation of estimates of the A and B-bases, requires additional computations; see Appendix C.1 of [3] for details.

Note that in contrast with the Bayesian approaches that have hitherto been suggested in this context, CVaR regression makes minimal assumptions about underlying distributions, and is therefore essentially nonparametric. Also, CVaR regression naturally pools together information from all other batches when predicting/fitting for any one particular batch; another feature absent from proposed Bayesian approaches.

The report is divided into three sections. Sections 10.2 and 10.3 are concerned with assessing the performance of CVaR regression in a controlled environment consisting of data simulated from known distributions. In Section 10.2 the model data is drawn from the same distribution as the test data, whilst in Section 10.3 it is drawn from a distribution different from the test data. Section 10.4 examines the performance of the methodology in the context of a supplied dataset similar to that analyzed in [3], augmented with failure data on two additional analytical models. Details on the implementation of the CVaR Regression methodology with an accompanying flowchart, are provided in the Appendix.

10.2 Assessing the Performance of CVaR Regression in a Controlled Environment: Model Data Drawn from Correct Distribution

In this section we investigate the performance of CVaR regression in a controlled environment consisting of data simulated from known Weibull distributions. The density function $f(x)$ of a two-parameter Weibull distribution with shape and scale parameters α and β respectively, is given by the formula,

$$f(x) = \frac{\alpha}{\beta}\left(\frac{x}{\beta}\right)^{\alpha-1} \exp\left\{-\left(\frac{x}{\beta}\right)^{\alpha}\right\}, \qquad \alpha > 0,\ \beta > 0.$$

Specifically, we consider the following problems.

- Calibration of the CVaR regression method by investigating the sensitivity to number of batches (m), and number of test points within each batch (n). As already stated, the term **batch** is used here in the broad sense to denote any one single source of variability affecting the test point data. For example: stacking sequence of laminates, type of test, batch number, etc. Any given batch is characterized by a particular choice of the shape and scale parameters of a Weibull distribution from which the failure load data for that batch is drawn.
- Relating CVaR deviation as defined by (10.2), to true error in 10th quantile estimation. The significance of this is immediately understood when we note that the former is observed while the latter is not.
- Quantifying the sensitivity of CVaR regression to errors/biases in the analytical model data.

We assess goodness-of-fit by measuring Mean Absolute Deviation (MAD) between true and estimated 10th quantiles across batches, and Mean CVaR Deviation (MCD) for CVaR regression fits. The CVaR deviation is that given by equation (10.2).

In this Section we consider only simulations where the model data is drawn from the same Weibull distribution as the test data; the latter viewed as the *correct* distribution. This "perfect information" setup allows us to benchmark the performance of the methodology in an idealized setting. By considering simulated model data drawn from a distribution that differs from that of the test data, Section 10.3 will address the problem discussed in the last point above.

Choice of Weibull model parameters

In order to simulate failure data that realistically corresponds to the values typically encountered in the supplied dataset (discussed in Section 10.4), we chose the range of the Weibull distribution shape and scale parameters as follows. The 19 batches that had at least 5 test points were selected. A Weibull distribution was then fitted via maximum likelihood to the failure values of all test points within each batch. Figure 10.1 shows the resulting estimated

parameters. Plots of the corresponding density functions appear in Figure 10.2. In all ensuing simulations in this Section, we draw failure loads randomly from Weibulls with shape and scale parameters in these ranges. Specifically, the ranges are:

$$10 \leq \alpha \leq 80, \qquad \text{and} \qquad 40 \leq \beta \leq 120.$$

Calibrating CVaR regression using only model data

We consider first CVaR regression using only model data as covariates (e.g. Model S). Plotted in the ensuing figures are the MAD and the MCD. The suffix "1" is appended to form MAD1 and MCD1, signaling that the corresponding deviations are based on CVaR regression fits using only **model** data as covariates. Other numbers will be appended later for other cases. The deviations are plotted as functions of number of batches (m) used in the CVaR regression fit, as well as number of test points (n) present in each batch.

Henceforth, and until further notice, a particular value of MAD or MCD is based on 100 draws from the same Weibull distribution. The choice of shape and scale parameters, constituting a batch, is made randomly from the region plotted in Figure 10.1. Each draw is of sample size $n + 100$. The first n points are then used as test data, and the remaining 100 as model data. A new set of shape and scale parameters is then randomly selected to produce data on the next batch, and so on. The formation of the CVaR regression inputs is done by using all test data as both response and covariates, as detailed in the Appendix. The deviations are plotted as functions of number of batches (m) used in the CVaR regression fit, as well as number of test points (n) within in each batch.

Figure 10.3, shows MAD1 error between true and estimated 10th quantiles for CVaR regression fits based on model data and different numbers (n) of test points, as a function of number of batches (m). Figure 10.4, shows the corresponding MCD1 error. Figure 10.5 shows part of the data from Figures 10.4 and 10.3 in the same plot. Generally, we see that MAD1 decreases while MCD1 increases, with increasing m. Also, MAD1 is lower while MCD1 is higher, at higher values of n, for any given m.

Figure 10.6, shows the MAD1 error between true and estimated 10th quantiles for CVaR regression fits based on model data and different numbers of test points (n) and batches (m), as a function of n. Figure 10.7, shows the corresponding MCD1 error. Figure 10.8 shows part of the data from Figures 10.7 and 10.6 in the same plot. Generally, we see that MAD1 decreases while MCD1 increases, with increasing n. Also, MAD1 is lower while MCD1 is higher, at higher values of m, for any given n.

Figures 10.9 and 10.10 provide a concise summary of the behavior of the MAD1 and MCD1 errors in Figures 10.3-10.8, via a three-dimensional perspective plot, as a function of both n and m. This offers a broader view than the two-dimensional plots above. Generally, MAD decreases with both m and n, levelling off at moderate values of these. MCD behaves inversely, increasing with m and n.

Summary Remarks:

- MAD1 decreases with increasing m (n), plateauing off at about $m = 10$ ($n = 10$). The rate of decrease is less pronounced at larger values of n (m).
- MCD1 increases with m (n), plateauing off at about $m = 10$ ($n = 10$). The rate of increase is less pronounced at larger values of n (m).
- For large m and n, MCD is between 4 to 7 times larger than MAD.

Calibrating CVaR regression using both model and test data

We now consider the case of model and test point data as covariates, and compute the mean absolute deviation (MAD3) between true and estimated 10th quantiles, and the mean CVaR

deviation (MCD3) for CVaR regression fits. The "3" signals that the corresponding deviation is based on CVaR regression fits using both **model** and **test** point data as covariates.

Figure 10.11, shows MAD3 error between true and estimated 10th quantiles for CVaR regression fits based on model data and different numbers (n) of test points, as a function of number of batches (m). Figure 10.12, shows the corresponding MCD3 error. Figure 10.13 shows part of the data from Figures 10.12 and 10.11 in the same plot. These graphs show that both MAD3 and MCD3 are relatively constant with m, the former (latter) being higher (lower) at lower values of n. As before, MCD3 is uniformly higher than MAD3, thus providing an upper bound on the MAD error.

Figure 10.14, shows the MAD3 error between true and estimated 10th quantiles for CVaR regression fits based on model data and different numbers of test points (n) and batches (m), as a function of n. Figure 10.15, shows the corresponding MCD3 error. Figure 10.16 shows part of the data from Figures 10.15 and 10.14 in the same plot. These graphs show that MAD3 (MCD3) decreases (increases) with n. Beyond $n = 5$, MCD3 is uniformly higher than MAD3, thus providing an upper bound on the MAD error.

Figures 10.17 and 10.18 provide a concise summary of the behavior of the MAD3 and MCD3 errors in Figures 10.11-10.16, via a three-dimensional perspective plot, as a function of both n and m. This offers a broader view than the two-dimensional plots above. Generally, MAD decreases with n, but is constant with m. MCD is also constant with m, but increases with n.

Summary Remarks:

- As a function of m, both MAD3 and MCD3 error are relatively constant, with lower (higher) values of MAD3 (MCD3) at higher (lower) values of n.
- MAD3 decreases from about 0.6 to 0.3 as n increases from 5 to 30.
- MCD3 increases from about 1 to 1.5 as n increases from 5 to 30.
- The decrease in MAD3 with increasing n, plateaus off at about $n = 10$.
- The increase in MCD3 with increasing n, plateaus off at about $n = 10$.
- For large m and n, MCD is between 2 to 6 times larger than MAD.

10.3 Assessing the Performance of CVaR Regression in a Controlled Environment: Model Data Drawn from Incorrect Distribution

To quantify the sensitivity of CVaR regression to errors in the model data, we consider deterministic and stochastic perturbations, with respect to test data (viewed as the *correct* distribution), in both the shape and scale parameters of the Weibull distribution from which the model data is drawn. The deterministic perturbations involve the changes

$$\alpha \to \alpha + 20, \quad \text{and} \quad \beta \to \beta + 20,$$

and the random perturbations the changes

$$\alpha \to \alpha + Z, \quad \text{and} \quad \beta \to \beta + Z.$$

(Z denotes a random value drawn from a Normal distribution with mean 0 and standard deviation 10.)

These perturbations can be thought of as introducing systematic and random biases in the model data.

Figure 10.19 shows MAD3 and MCD3 error, by number of test points (n), as a function of number of batches (m), as in Figure 10.13. Column 2 (3) shows the effect of deterministic and

random perturbations in the shape (scale) parameter. Figure 10.13 (no bias) is reproduced in the first column for comparison. These graphs show that MCD3 is virtually unaffected by both shape and scale biases. MAD3 is moderately affected by scale biases, increasing dramatically at lower values of m and n.

Figure 10.20 shows MAD3 and MCD3 error, by number of batches (m), as a function of number of test points (n), as in Figure 10.16. Column 2 (3) shows the effect of deterministic and random perturbations in the shape (scale) parameter. Figure 10.16 (no bias) is reproduced in the first column for comparison purposes. Again, MCD3 is virtually unaffected by both shape and scale biases. MAD3 is severely affected by scale biases, increasing dramatically at lower values of m and n.

Figures 10.21-10.28 show MAD3 and MCD3 error perspective plots similar to Figures 10.17 and 10.18, but with deterministic and random shape and scale perturbations. Figure 10.29 (Figure 10.30) combines all the MAD3 (MCD3) plots into one, with Figure 10.17 (Figure 10.18) reproduced in the first column for comparison. MAD3 decreases with m and n as already noted, with a dramatic increase at low values of m and n with respect to the unbiased plots. The plots also clearly show that MCD3 is relatively unaffected by both shape and scale biases.

Summary Remarks:

- MAD is virtually unaffected by shape biases.
- MAD is moderately affected by scale biases; dramatically at lower values of m and n.
- MCD is virtually unaffected by both shape and scale biases.

The fact that the location of a Weibull depends more on its scale parameter β than on its shape parameter α, is undoubtedly partly responsible for the larger impact of scale biases. Another way to assess this is to consider the Kullback-Leibler distance between the perturbed Weibull and the true one. The Kullback-Leibler distance of distribution g from distribution f, is a measure of the *information lost when g is used to approximate f*; Kullback and Leibler (1951). Figure 10.31 plots the Kullback-Leibler distance of various Weibulls from a Weibull(40,80), as a function of the approximating Weibull's shape and scale parameters. As we can see, this distance varies more with scale than it does with shape, and rises dramatically when the approximating Weibull's scale parameter is too low.

10.4 Assessing the Predictive Capabilities of CVaR Regression with Data from Several Analytical Models

In this section we apply the CVaR regression methodology to the supplied dataset similar to the *full dataset* of [3], augmented with failure load information from two additional analytical models. From a total of 28 batches, only the 19 batches with at least 5 test points per batch were selected. As in [3], the dataset was symmetrized by selecting exactly 5 test points per batch. In addition to the Model S failure load prediction model data within each batch already present, data on two additional models, Model S2 and Model S3, was added. The objective is to extend the results in [3], thus quantifying the benefits of introducing other models.

We repeat Analyses 1 and 2 of [3]. The 10th percentile estimated surface for a generic factor model under consideration, takes on the form:

$$\hat{Q}_Y(0.10) = \hat{c}_0 + \hat{c}_1 m + \hat{c}_2 s + \hat{c}_3 \mu^{(1)} + \hat{c}_4 \sigma^{(1)} + \hat{c}_5 \mu^{(2)} + \hat{c}_6 \sigma^{(2)} + \hat{c}_7 \mu^{(3)} + \hat{c}_8 \sigma^{(3)},$$

where the superscript (i), $i = 1, 2, 3$, identifies the Model used to compute the summary statistic corresponding to that particular coefficient. The method of using complementary subsets

Table 10.1. Results of CVaR regression applied to the supplied dataset with test points ($\hat{c}_1$,$\hat{c}_2$) and the three Models: Model S ($\hat{c}_3$,$\hat{c}_4$), Model S2 ($\hat{c}_5$,$\hat{c}_6$) and Model S3 ($\hat{c}_7$,$\hat{c}_8$). Within each batch, all 5 test points were used as output, but only a subset was used as input (column 1).

Number of	Estimated Coefficients of $Q_Y(0.10)$									Goodness-of-fit
Test Points	Const.	Test Data		Model S		Model S2		Model S3		CVaR Measure
Used as Input	$\hat{c}_0$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	$\hat{c}_4$	$\hat{c}_5$	$\hat{c}_6$	$\hat{c}_7$	$\hat{c}_8$	$(\mathcal{P}^2_{0.10})$
0	36.73							0.59	-0.31	27.16
0	39.26					0.57	0.01			24.66
0	14.27			1.10	-4.30					16.86
0	13.32			0.51	-6.01	0.66	-1.24	0.04	0.04	13.82
1	14.29	0.30		0.11	-5.13	0.83	-1.06	-0.08	0.07	12.61
2	15.75	0.44	0.22	-0.26	-5.71	1.16	-1.03	-0.18	0.09	12.36
3	12.75	0.62	-0.27	-0.14	-3.88	0.71	-0.72	-0.09	0.05	11.82
4	11.08	0.88	-0.88	-0.10	-1.64	0.33	-0.37	-0.06	0.03	10.79
5	10.30	1.00	-1.44							10.29
5	10.88	0.94	-1.46	0.07	-0.26					10.21
5	11.30	1.04	-1.44					-0.04	0.03	10.14
5	11.00	1.01	-1.43			0.01	-0.22			10.01
5	9.27	0.97	-1.43	0.16	0.16	-0.11	-0.18	-0.00	0.04	9.72

for response and covariate CVaR regression input formation, outlined in the Appendix, is used here.

The results are summarized in Table 10.1. The form of the table allows us to easily identify how many test points and which Models were used as input in which case. For example, in row 2 no test points and only Model S2 were used as input. The results are also ranked according to the CVaR goodness-of-fit measure, $\mathcal{P}^2_{0.10}$. According to this criterion, we can conclude the following concerning these various sources of input to CVaR regression:

- Among the Models used as the sole input with no test data, the best performer is Model S, followed by Model S2. Model S3 is the worst.
- With no test data as input, the contribution from all 3 analytical models results in an improvement of approximately:
  - 50% over Model S3 by itself;
  - 44% over Model S2 by itself;
  - 18% over Model S by itself.

  The combining of model information is therefore advantageous.
- When all 5 test points are used as input, each of the 3 Models by themselves provides practically no additional improvement (less than 3%). The addition of all 3 Models results in only about a 6% improvement.
- When all 5 test points are used as input, each of the 3 Models by themselves provides practically no additional improvement (less than 3%). The addition of all 3 Models results in only about a 6% improvement.
- The models contribute substantially to percentile prediction when up to 3 test points are used as input. This can be seen in Table 10.1 as the magnitude of the estimated coefficients shifts from the model to the test data as the number of test points used as input increases.
- The contribution from 4 test points plus all 3 Models, is approximately equivalent to the contribution from 5 test points by themselves. This implies the 3 Models can be roughly equated to one test point.

Figure 10.32 gives a visual summary of the rate of decrease in $\mathcal{P}^2_{0.10}$ as a function of the fitted model (row number in Table 10.1).

We have noticed that the reason for the under-performance of Models 2 and (especially) 3 relative to Model 1, is that they have more outliers with respect to the test data (the range of

failure loads predicted by these Models within batches often fails to capture any test data). This suggests the need for an additional calibration procedure before using a given analytical model. We performed one such crude calibration, by removing from the data batches whose range of values predicted by Model 1 failed to capture any test point data. Only 12 batches remained. The same was done for Models 2 and 3, but the criterion for batch exclusion there was that there should be no test points within one standard deviation of the corresponding Model mean. Ten (10) and 5 batches remained, respectively. CVaR regression was run on only test points plus each model. The results are summarized in Table 10.2. Note that the values of

Table 10.2. Results of CVaR regression applied to the supplied dataset with outlying batches removed. The coefficients correspond to test points ($\hat{c}_1$,$\hat{c}_2$) and the three Models: Model S ($\hat{c}_3$,$\hat{c}_4$), Model S2 ($\hat{c}_5$,$\hat{c}_6$) and Model S3 ($\hat{c}_7$,$\hat{c}_8$). Within each batch, all 5 test points were used as output and their summary statistics used as input. Batches were excluded if the Model failed to capture any of its test points (Model S), or if there were no test points within one standard deviation of the Model mean (Model S2 and Model S3).

Number	Estimated Coefficients of $Q_Y(0.10)$									$\mathcal{P}^2_{0.10}$
of	Const.	Test Data		Model S		Model S2		Model S3		Table 10.2
Batches	$\hat{c}_0$	$\hat{c}_1$	$\hat{c}_2$	$\hat{c}_3$	$\hat{c}_4$	$\hat{c}_5$	$\hat{c}_6$	$\hat{c}_7$	$\hat{c}_8$	(Table 10.1)
12	0.24	0.98	-1.55	0.02	0.02					10.04 (10.21)
10	2.27	1.16	-1.25			-0.18	-0.27			9.94 (10.01)
5	-3.41	1.26	-1.32					-0.22	0.19	7.24 (10.14)

the CVaR goodness-of-fit measure $\mathcal{P}^2_{0.10}$ have all decreased from their values in Table 10.1 (in parentheses in last column). The reason for this is in part due to an improved fit, and in part due to a reduction in the number of batches used in model fitting.

10.5 Suggestions for Future Research

The simulations and case studies presented thus far show that ours is a promising methodology for integrating various sources of variability on properties of materials. However, only the simplest of scenarios have hitherto been tackled; the methodology clearly warrants further and deeper investigation. In this section we outline some possible directions for future research.

1. We have hitherto assumed that the experimental test data is observed without error, i.e. it comes from the true failure distribution. The extent to which CVaR regression estimates are impacted when the test data is contaminated with errors, deserves careful attention.
2. Coefficient estimation in the current form of CVaR regression is wholly data driven, in the sense that there is no prior weighting of factors to reflect the suspected precision of a particular factor relative to the others. In ordinary regression, this is achieved by choosing appropriate weights for least squares parameter estimation, giving rise to *weighted regression*. Analogous ideas could be applied to CVaR regression.
3. The performance of the methodology in its present form, has only been assessed on 10th percentile inference (B-basis). Inference on the 1st percentile (A-basis), may require quite a large number of test data points per batch in order to be effective. Specialized variants of CVaR regression may need to be developed to cope with this situation.
4. As a follow-up to the previous point, it may be possible to adapt some form of Extreme Value Theory (EVT) to provide an alternative to, or be used in conjunction with, CVaR regression. EVT methods were designed for inference on the tails of distributions where data is usually sparse. It may be possible to adapt some of its key ideas in order to alleviate the problem of insufficient test data when A-basis (or even B-basis) estimates are desired.

5. Assess the capabilities of the methodology for pooling experimental data from various sources, and compare this with Bayes and empirical Bayes methods.
6. Use a Bayesian hierarchical approach instead of CVaR regression to obtain the predictive distribution of the failure load. The contribution from each analytical model could be assessed/incorporated via a Bayesian hyperprior of weights.
7. The suggested CVaR methodology is nonparametric. Assess the impact of parametric assumptions on distributions on the Bayesian approach, in order to obtain estimates of similar quality to CVaR regression.

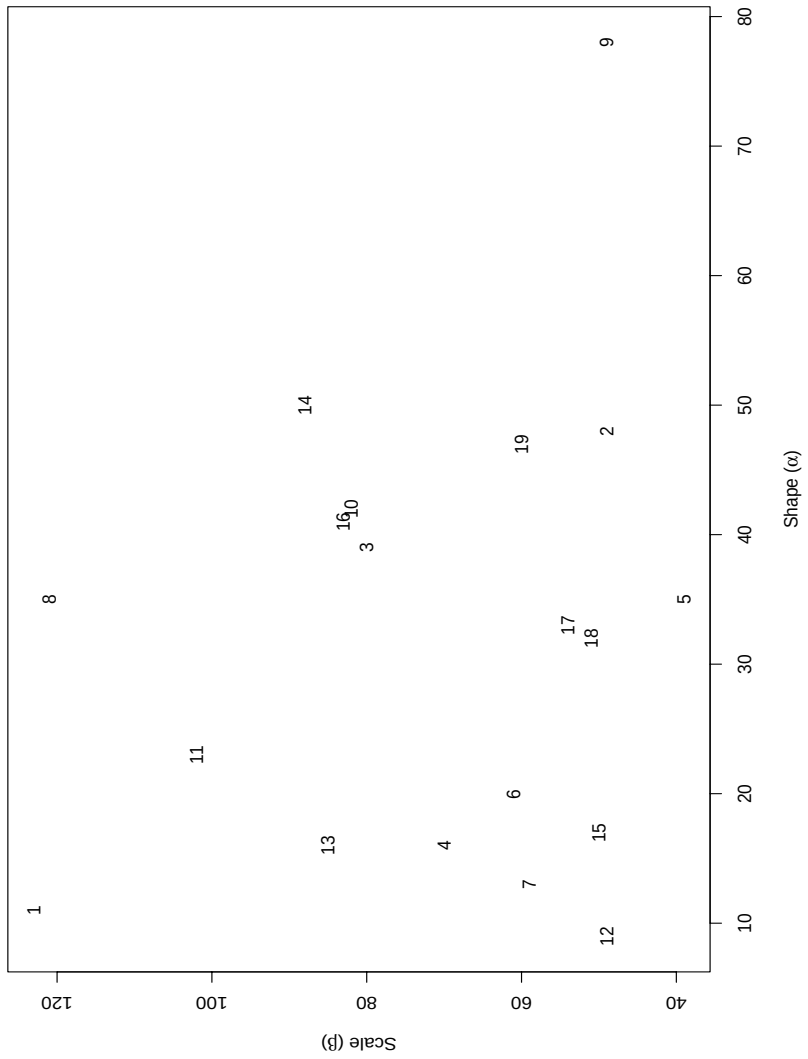


Fig. 10.1. Estimated shape and scale parameters of a Weibull fit to the test data from the 19 batches of the supplied dataset that had at least 5 test points per batch. Plots of the corresponding density functions appear in Figure 10.2.

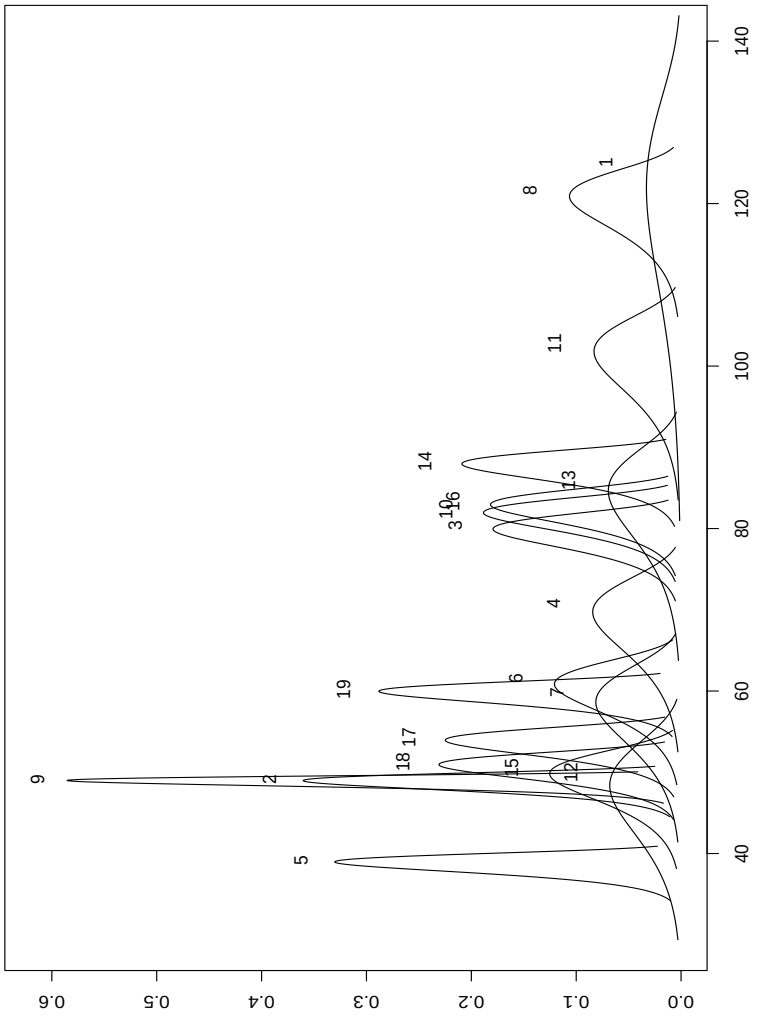


Fig. 10.2. Plots of the Weibull density functions corresponding to the shape and scale parameters appearing in Figure 10.1.

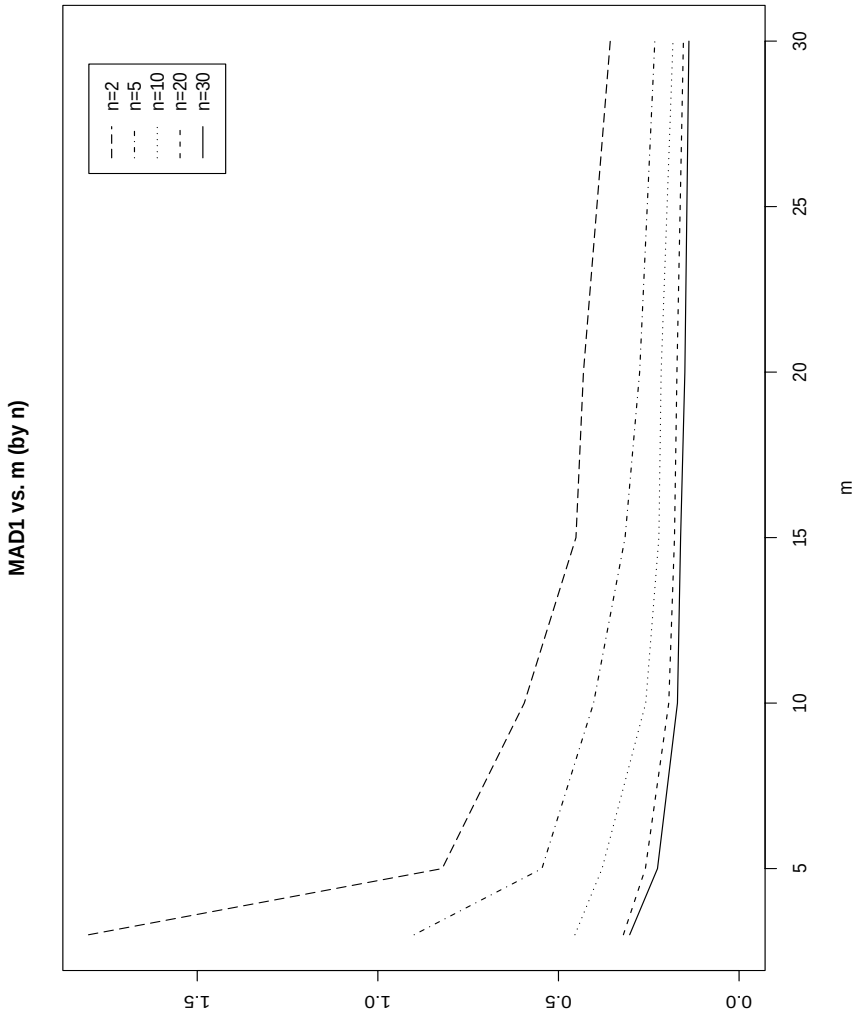


Fig. 10.3. Mean Absolute Deviation (MAD1) error between true and estimated 10th quantiles for CVaR regression fits based on model data and different numbers of test points (n), as a function of number of batches (m). Only model data was used as covariates. MAD1 is based on 100 draws from the same Weibull distribution within each batch, each draw of sample size $n + 100$ (n used as test data; 100 as model data).

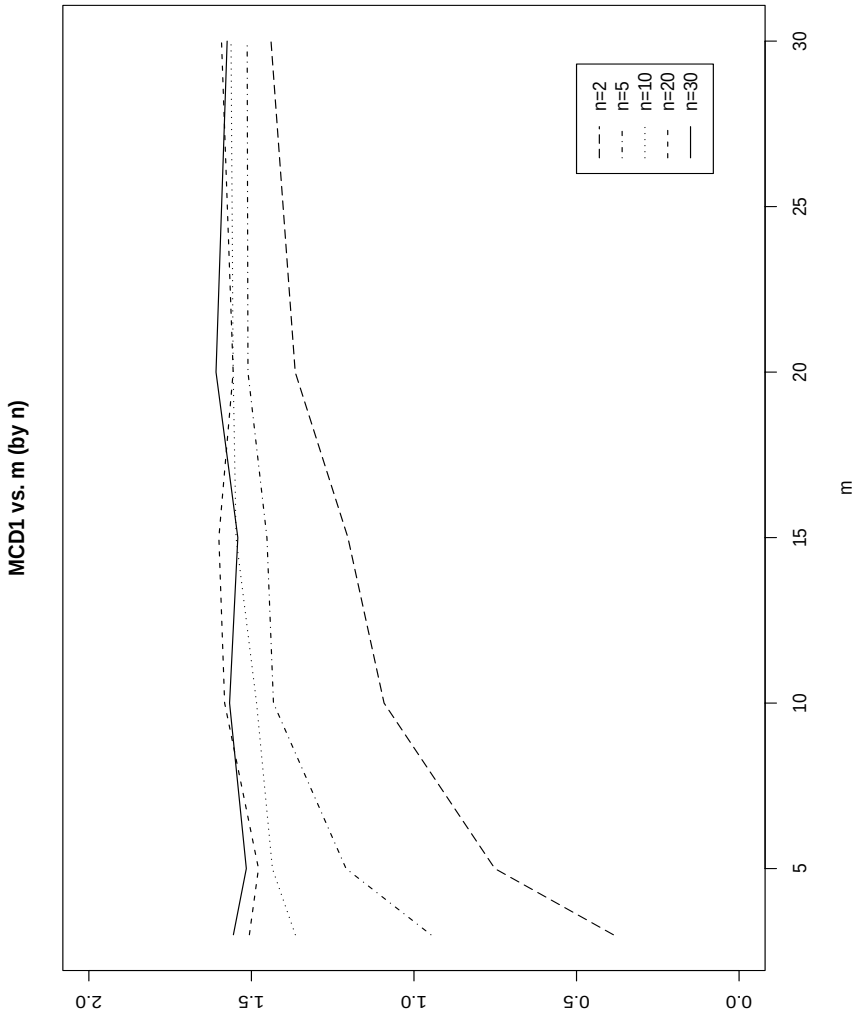


Fig. 10.4. Mean CVaR Deviation (MCD1) error for CVaR regression fits based on model data and different numbers (n) of test points, as a function of number of batches (m). Only model data was used as covariates. MCD1 is based on 100 draws from the same Weibull distribution within each batch, each draw of sample size $n + 100$ (n used as test data; 100 as model data.

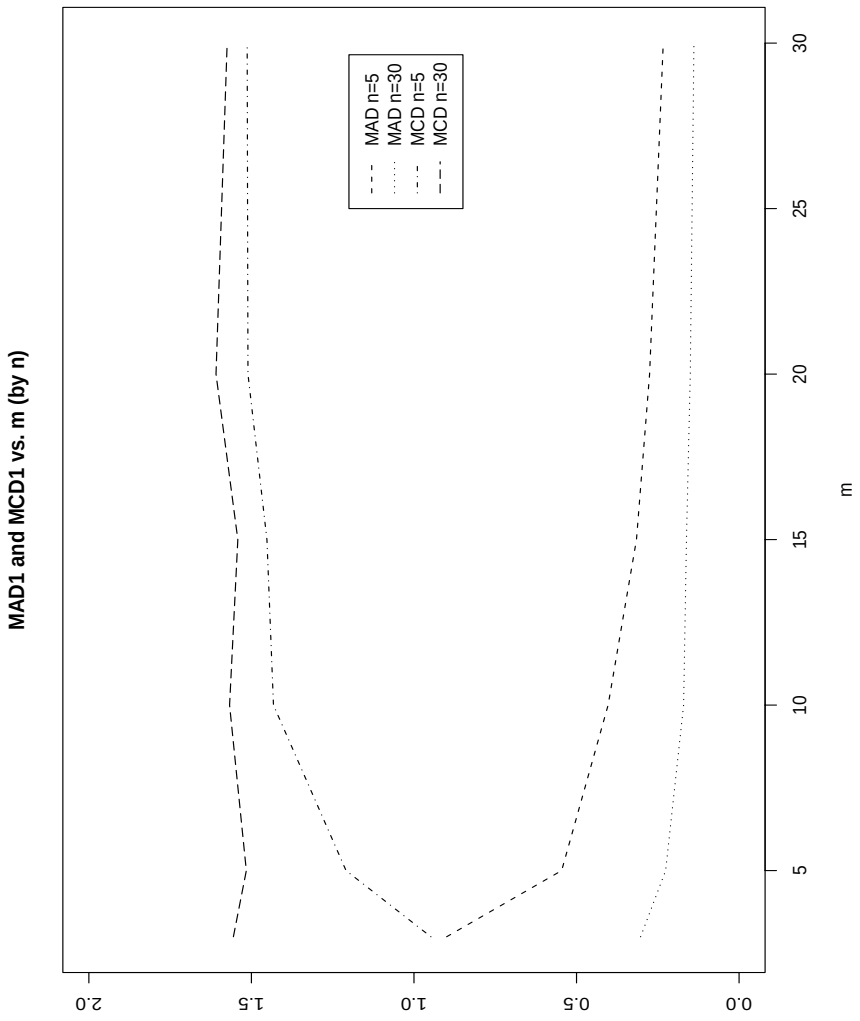


Fig. 10.5. MAD1 and MCD1 error from Figures 10.4 and 10.3 combined in the same plot.

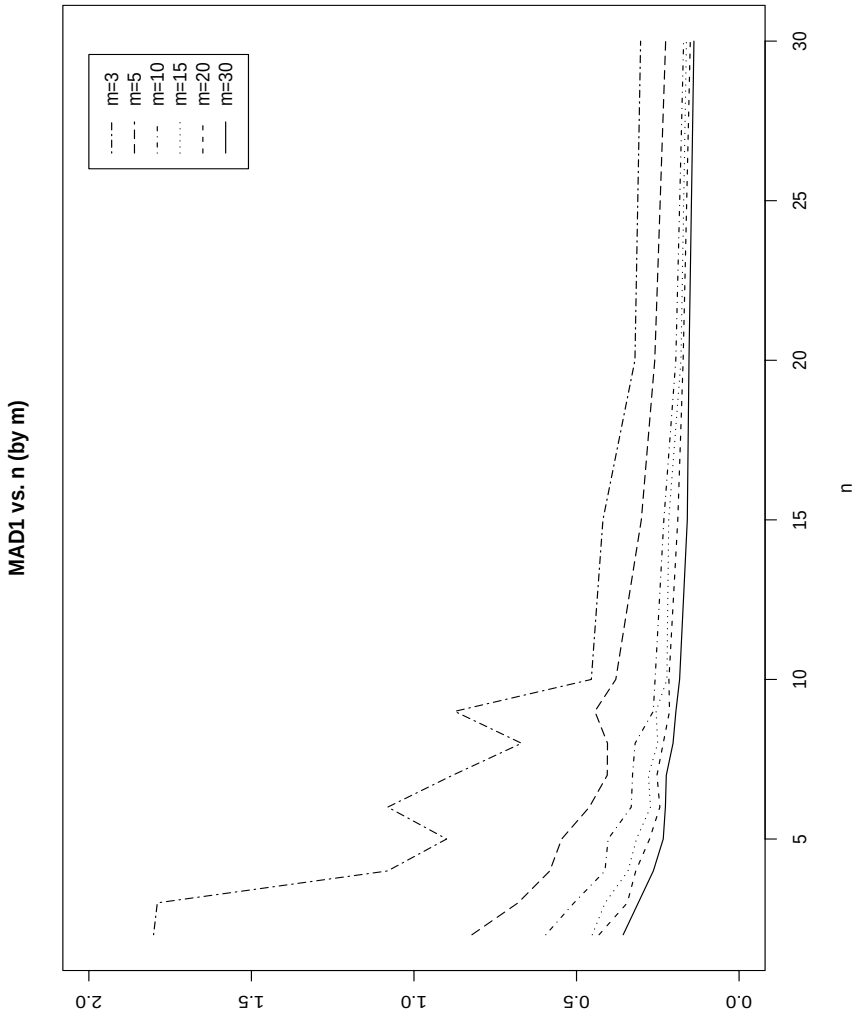


Fig. 10.6. Mean Absolute Deviation (MAD1) error between true and estimated 10th quantiles for CVaR regression fits based on model data and different numbers of test points (n) and batches (m), as a function of n.

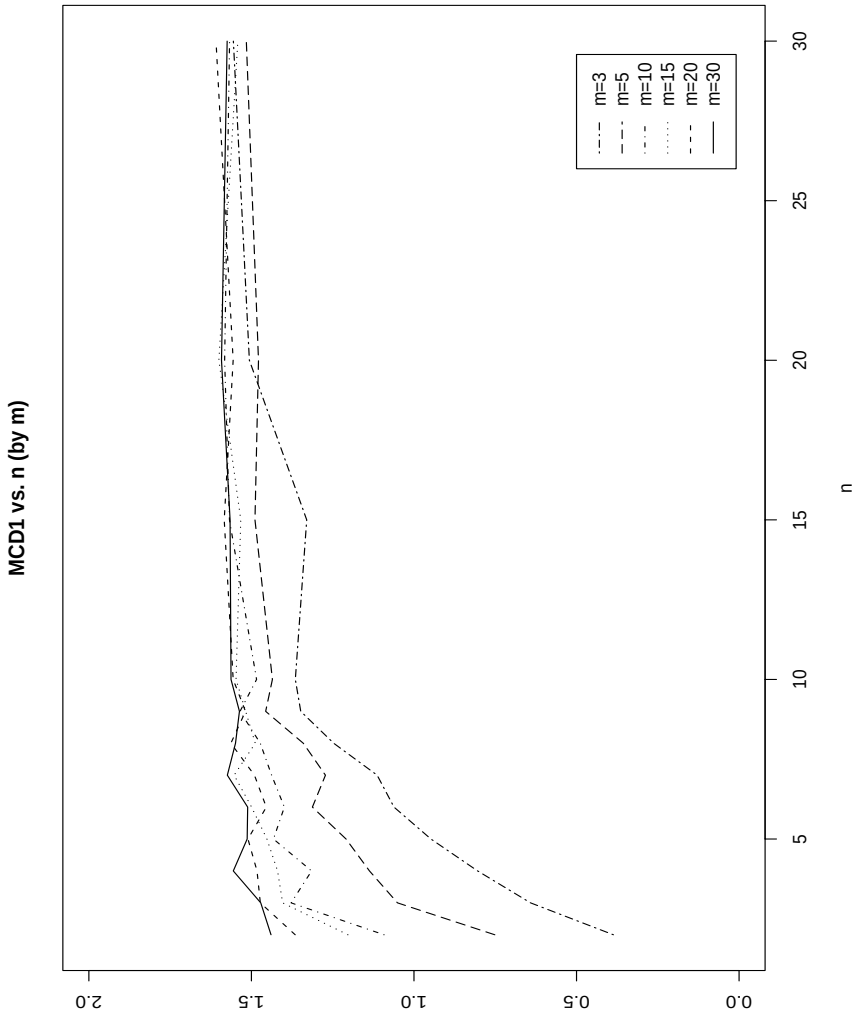


Fig. 10.7. Mean CVaR Deviation (MCD1) error between true and estimated 10th quantiles for CVaR regression fits based on model data and different numbers of test points (n) and batches (m), as a function of n.

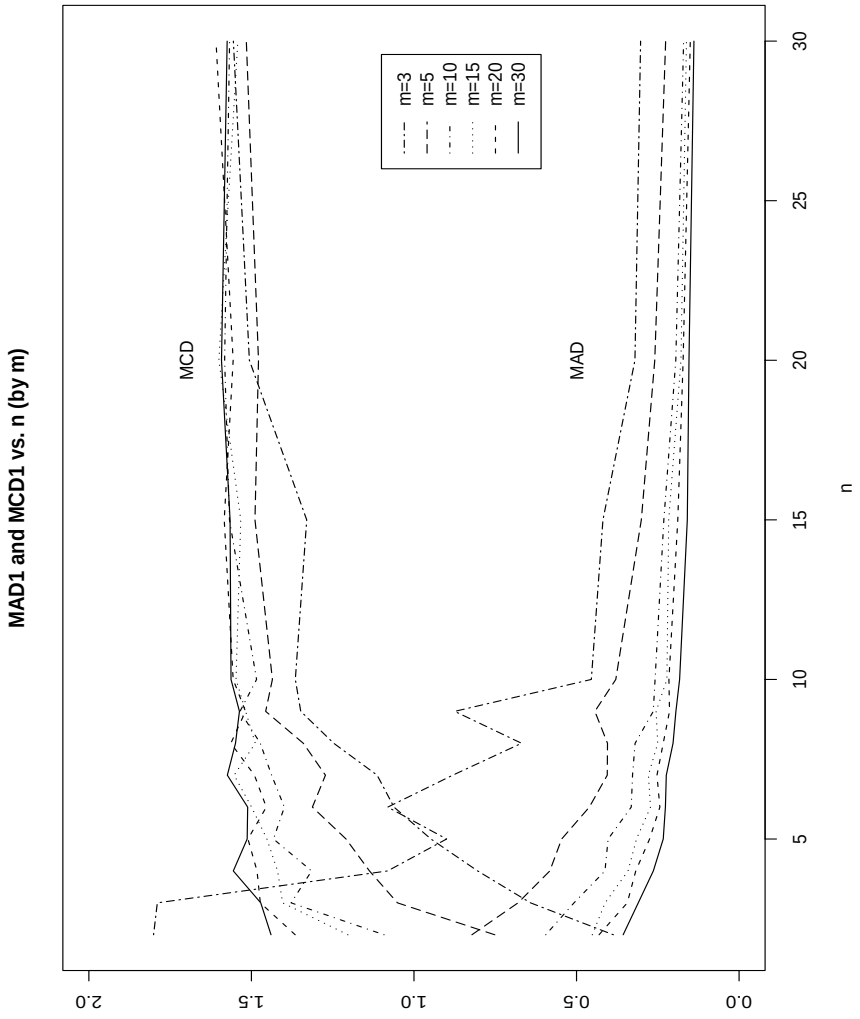


Fig. 10.8. MAD1 and MCD1 error from Figures 10.7 and 10.6 combined in the same plot.

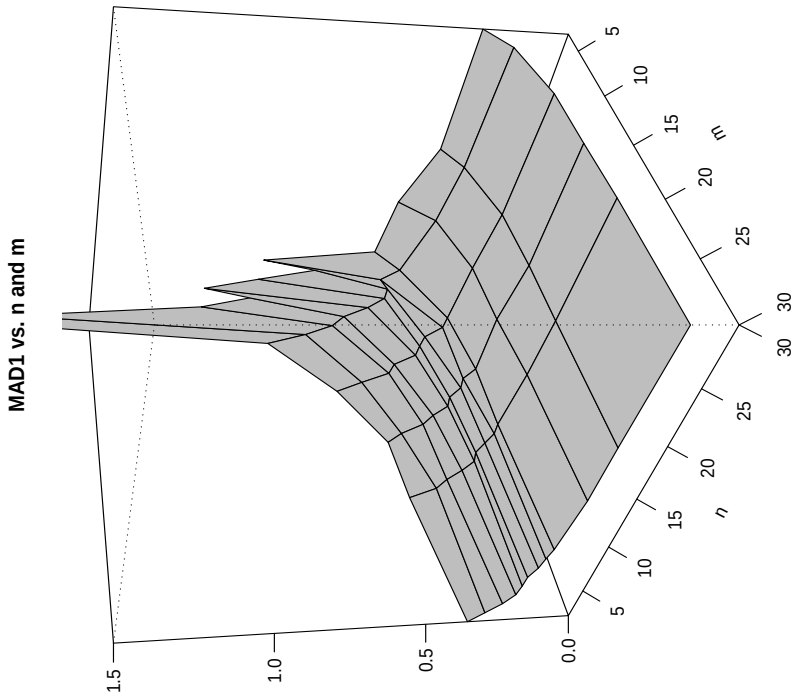


Fig. 10.9. MAD1 error perspective plot of the data in Figures 10.3-10.7, as a function of both number of test points (n) and number of batches (m).

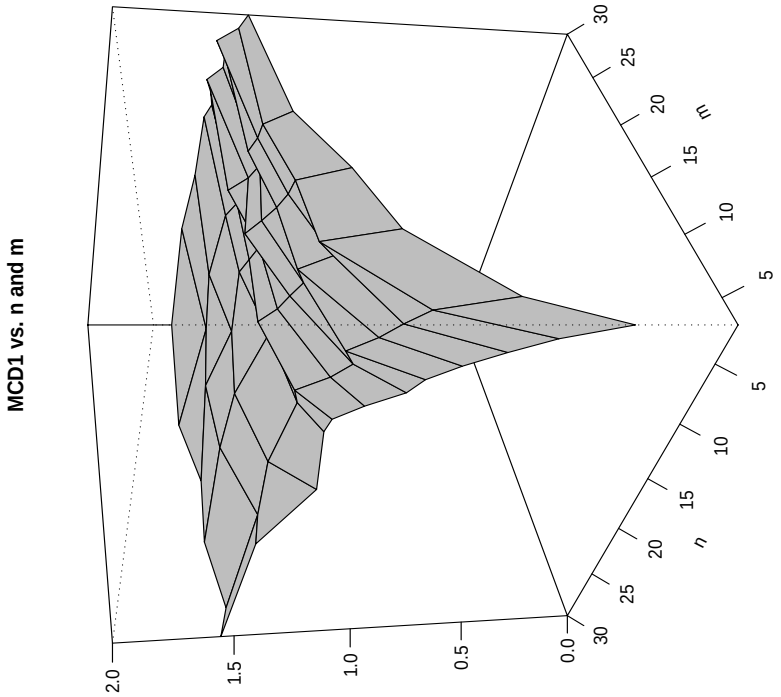


Fig. 10.10. MCD1 error perspective plot of the data in Figures 10.3-10.7, as a function of both number of test points (n) and number of batches (m).

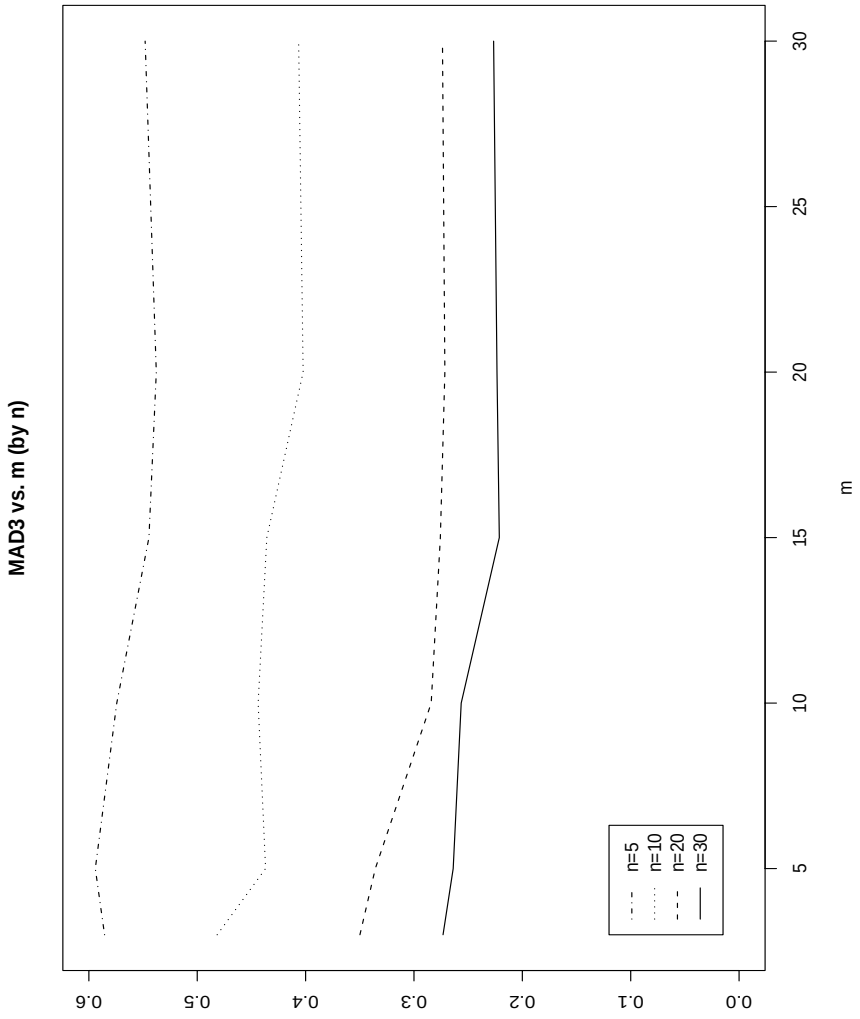


Fig. 10.11. Mean Absolute Deviation (MAD3) error between true and estimated 10th quantiles for CVaR regression fits based on model data and different numbers of test points (n), as a function of number of batches (m). Both model and test data were used as covariates. MAD3 is based on 100 draws from the same Weibull distribution within each batch, each draw of sample size $n + 100$ (n used as test data; 100 as model data).

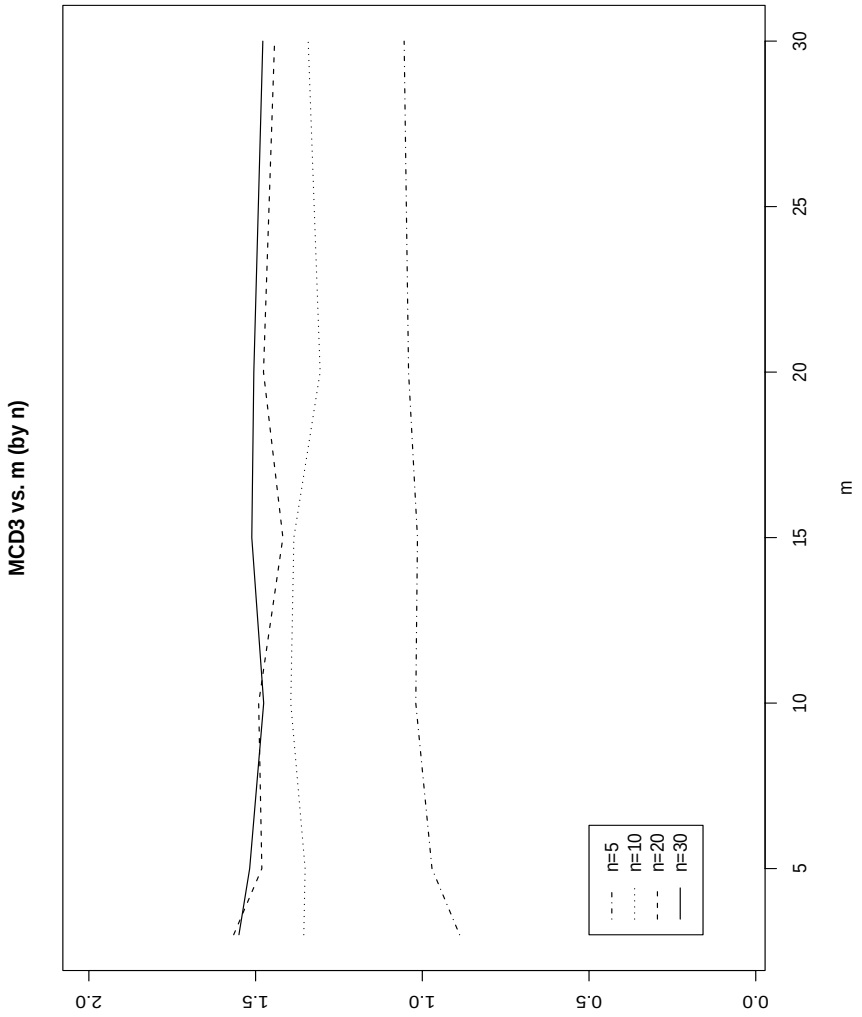


Fig. 10.12. Mean CVaR Deviation (MCD3) error for CVaR regression fits based on model data and different numbers (n) of test points, as a function of number of batches (m). Both model and test data were used as covariates. MCD3 is based on 100 draws from the same Weibull distribution within each batch, each draw of sample size $n + 100$ (n used as test data; 100 as model data.

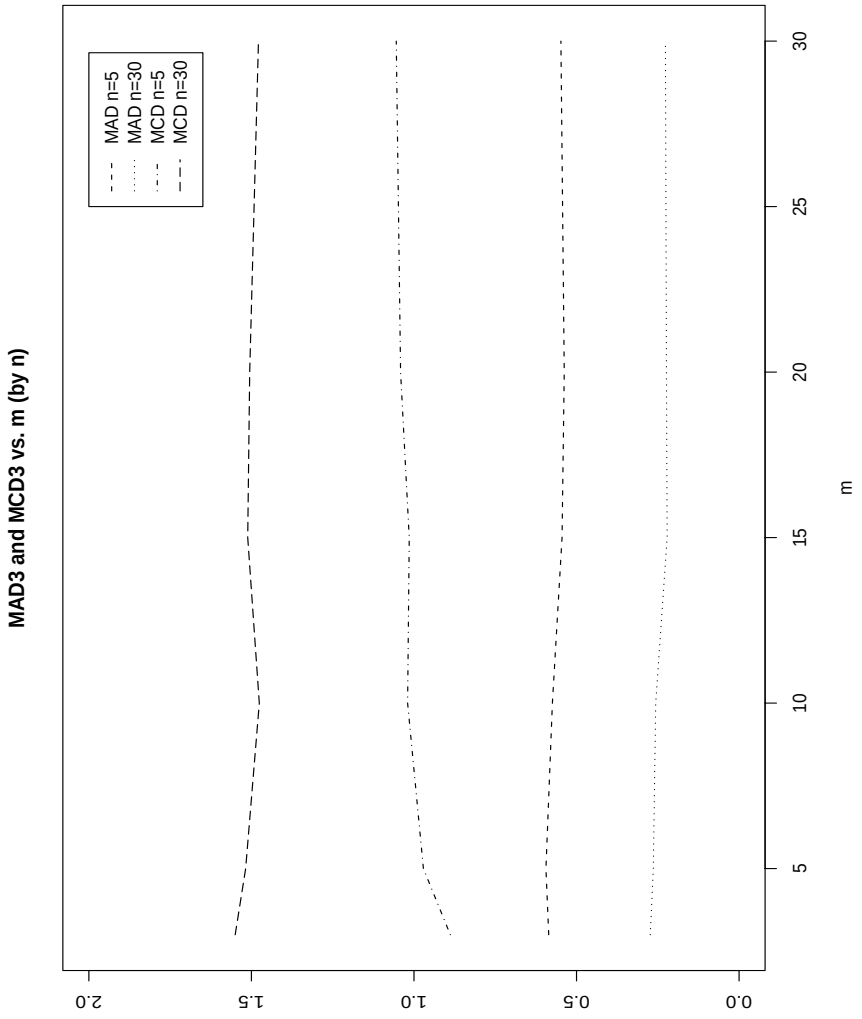


Fig. 10.13. MAD3 and MCD3 error from Figures 10.12 and 10.11 combined in the same plot.

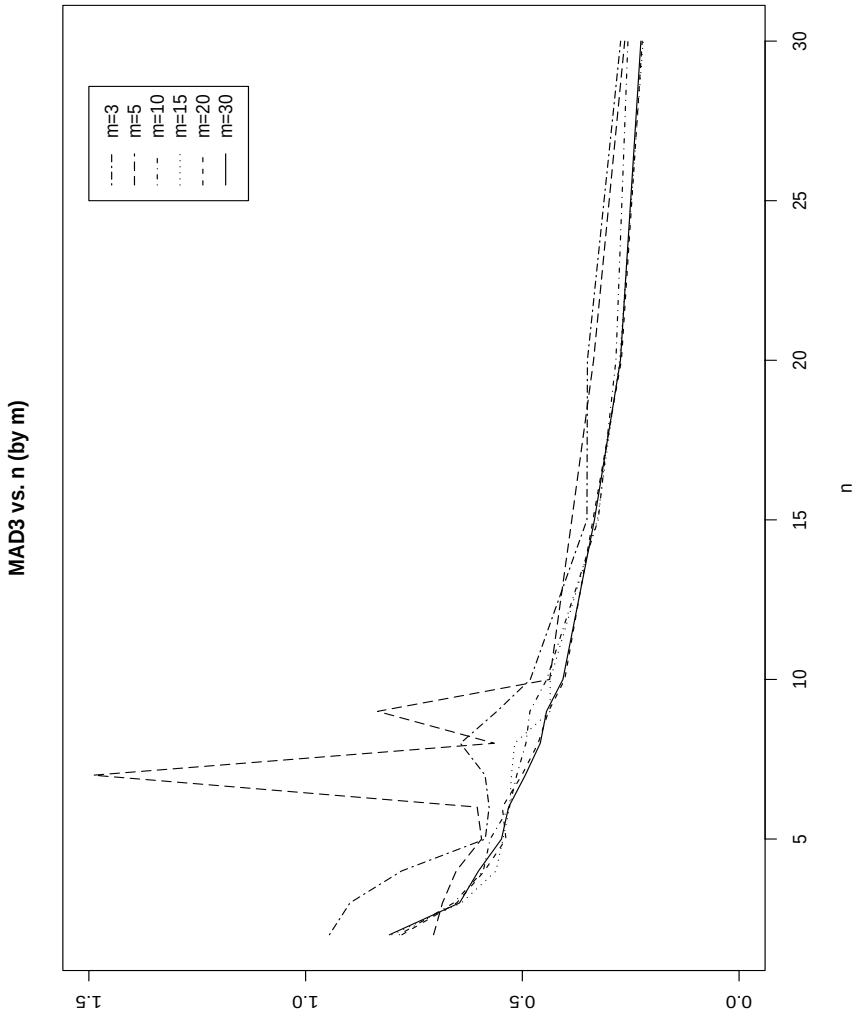


Fig. 10.14. Mean Absolute Deviation (MAD3) error between true and estimated 10th quantiles for CVaR regression fits based on model data and different numbers of test points (n) and batches (m), as a function of n.

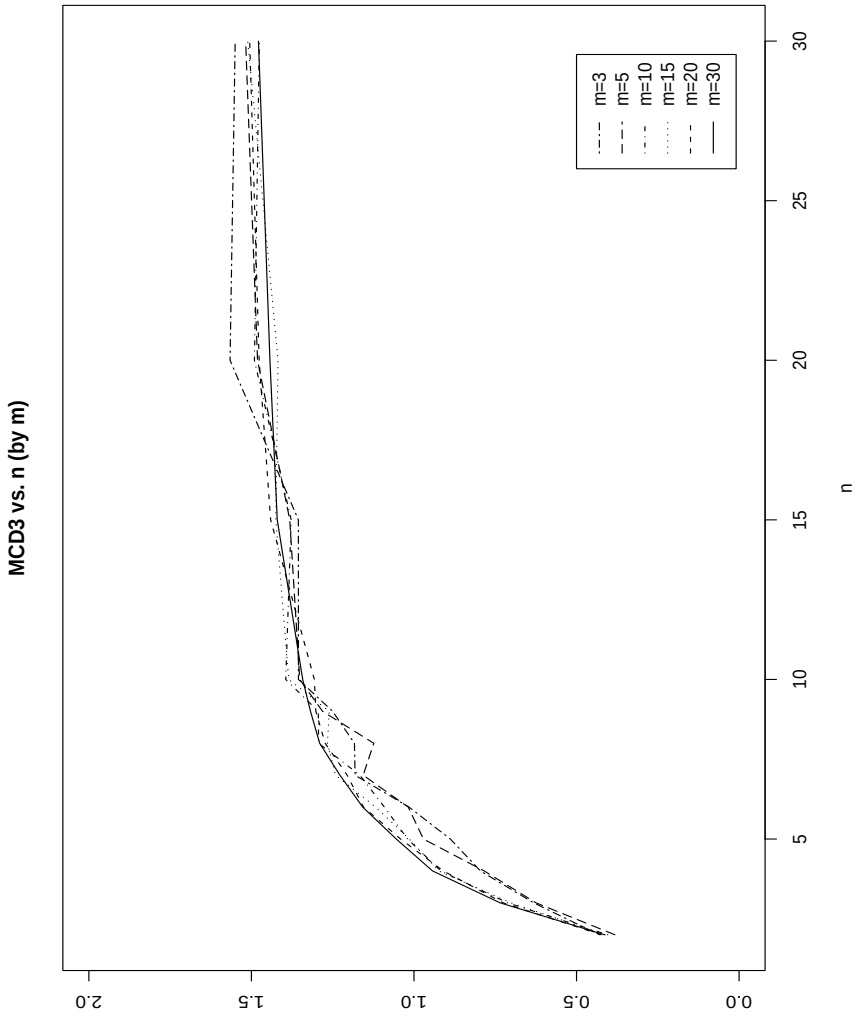


Fig. 10.15. Mean CVaR Deviation (MCD3) error between true and estimated 10th quantiles for CVaR regression fits based on model data and different numbers of test points (n) and batches (m), as a function of n.

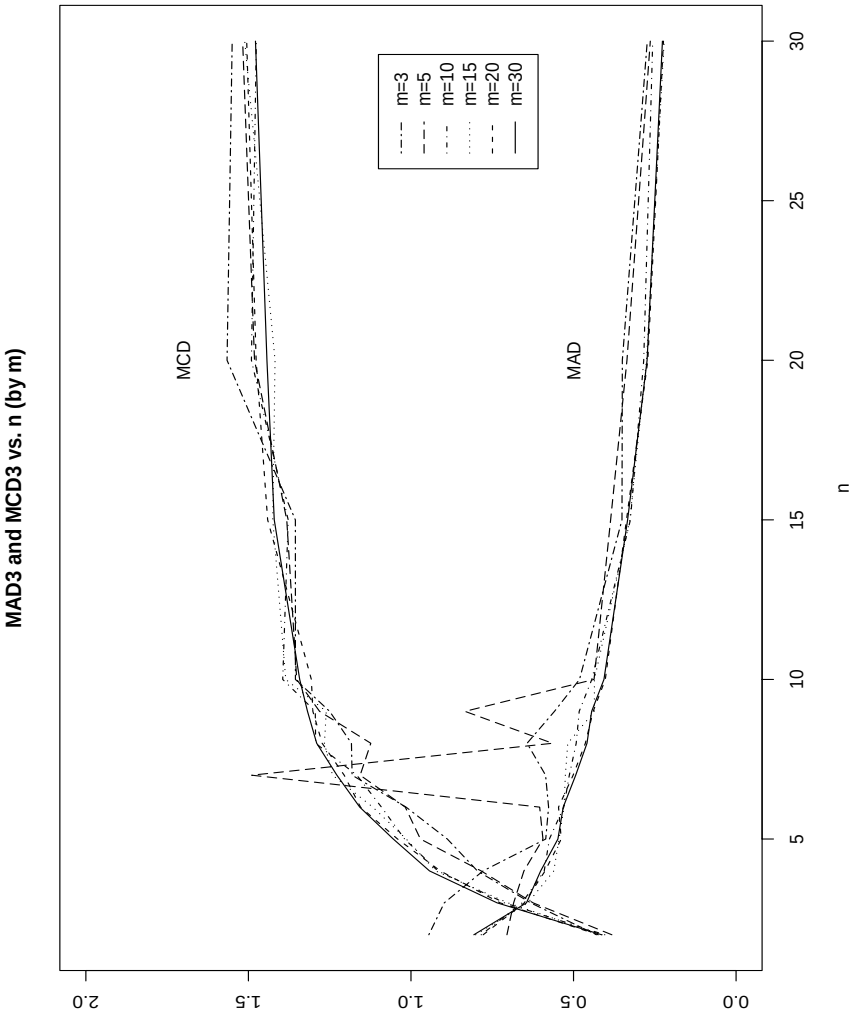


Fig. 10.16. MAD3 and MCD3 error from Figures 10.15 and 10.14 combined in the same plot.

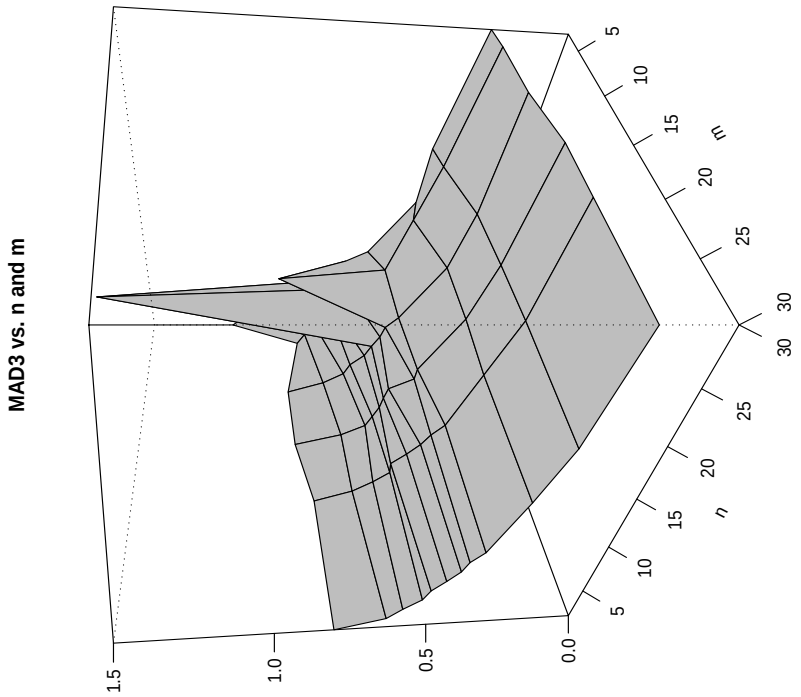


Fig. 10.17. MAD3 error perspective plot of the data in Figures 10.11-10.15, as a function of both number of test points (n) and number of batches (m).

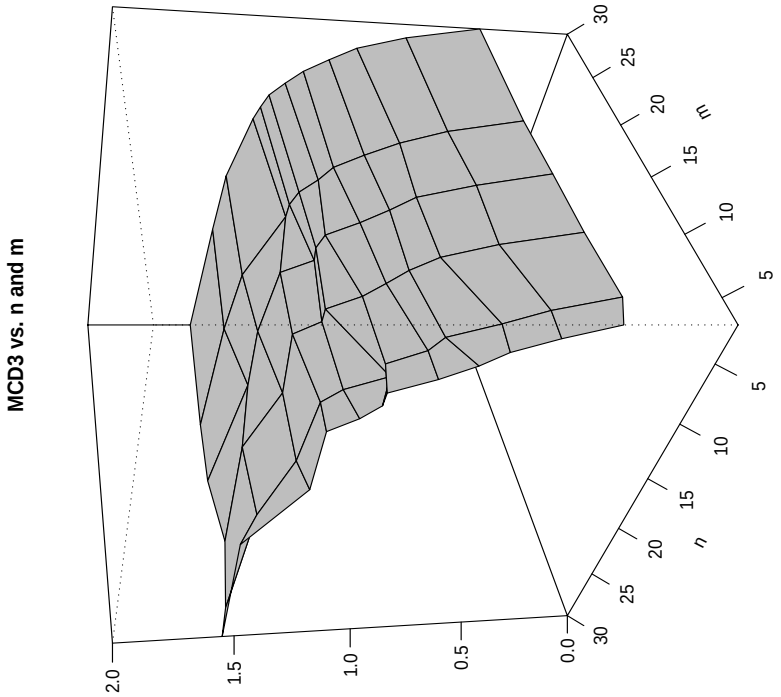


Fig. 10.18. MCD3 error perspective plot of the data in Figures 10.11-10.15, as a function of both number of test points (n) and number of batches (m).

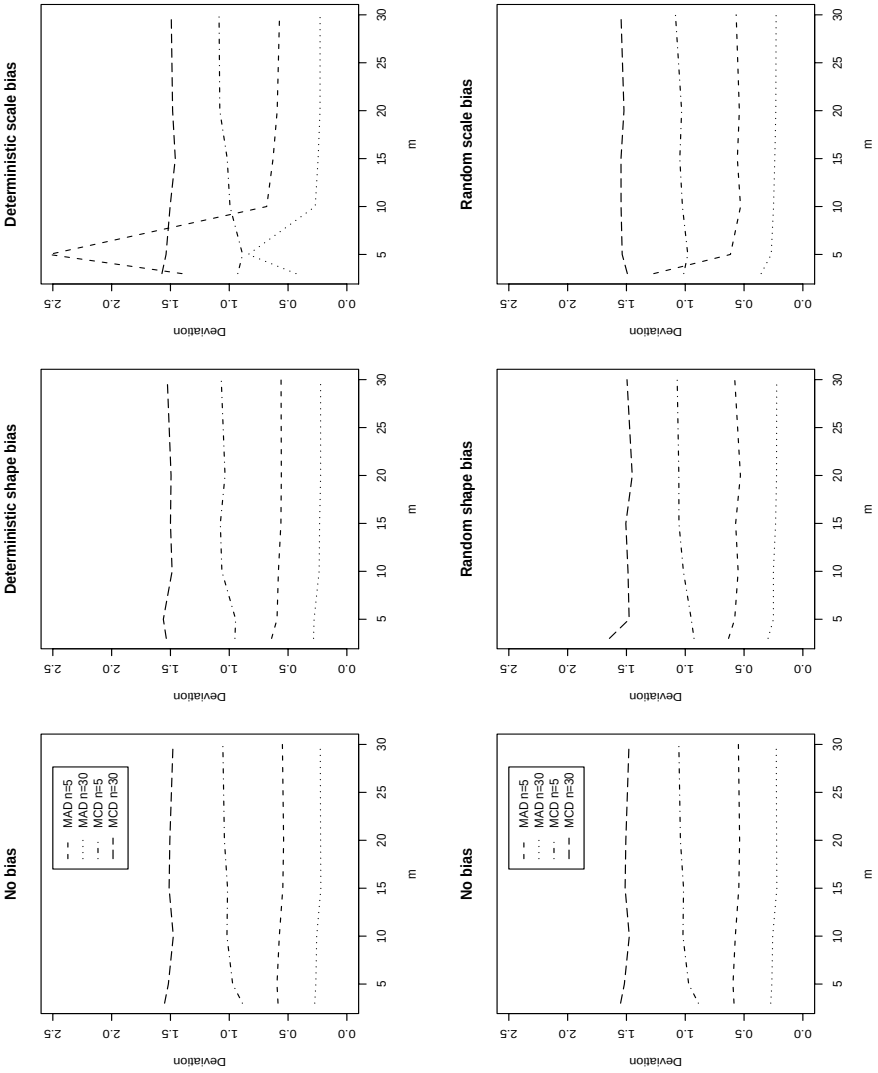


Fig. 10.19. MAD3 and MCD3 errors similar to Figure 10.13 (no bias, column 1), but with deterministic perturbations in the shape ($\alpha \to \alpha + 20$, column 2) and scale ($\beta \to \beta + 20$, column 3) parameters of the Weibull distribution from which the model data is drawn.

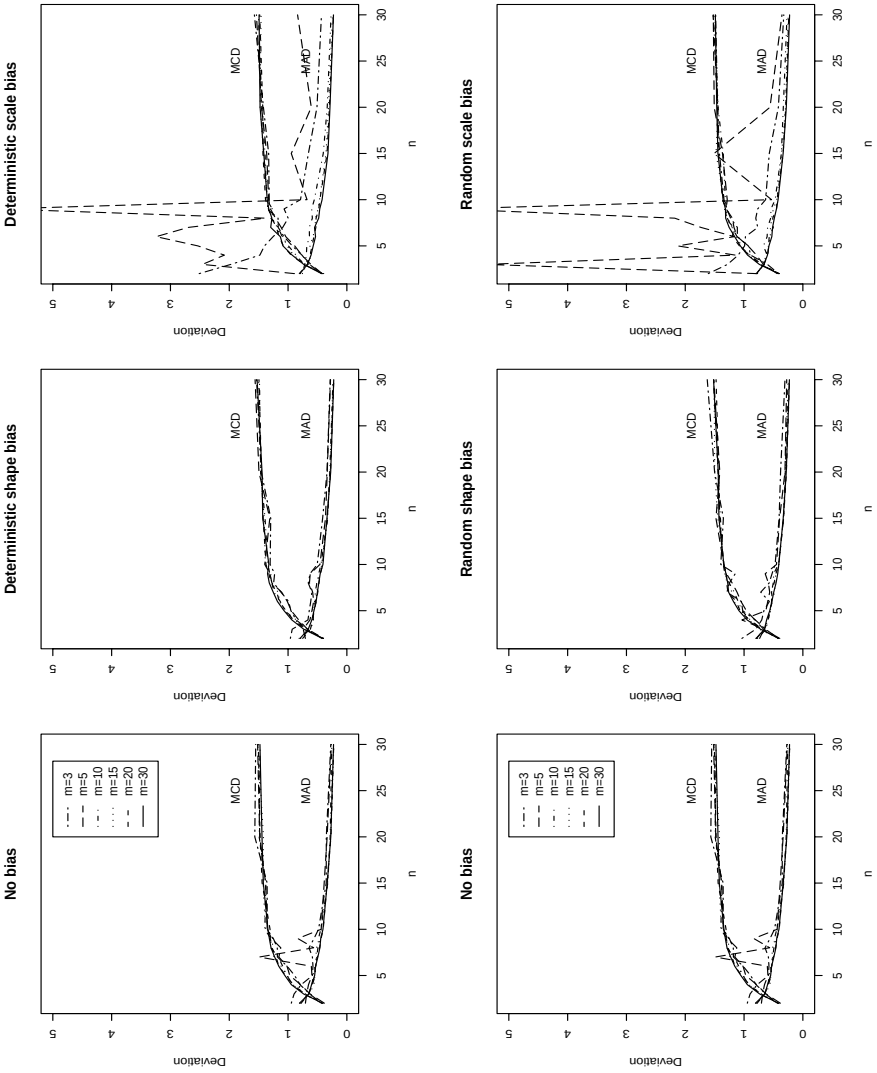


Fig. 10.20. MAD3 and MCD3 errors similar to Figure 10.16 (no bias, column 1), but with random perturbations in the shape ($\alpha \rightarrow \alpha + Z$, column 2) and scale ($\beta \rightarrow \beta + Z$, column 3) parameters of the Weibull distribution from which the model data is drawn. (Z denotes a random value drawn from a Normal distribution with mean 0 and standard deviation 10.)

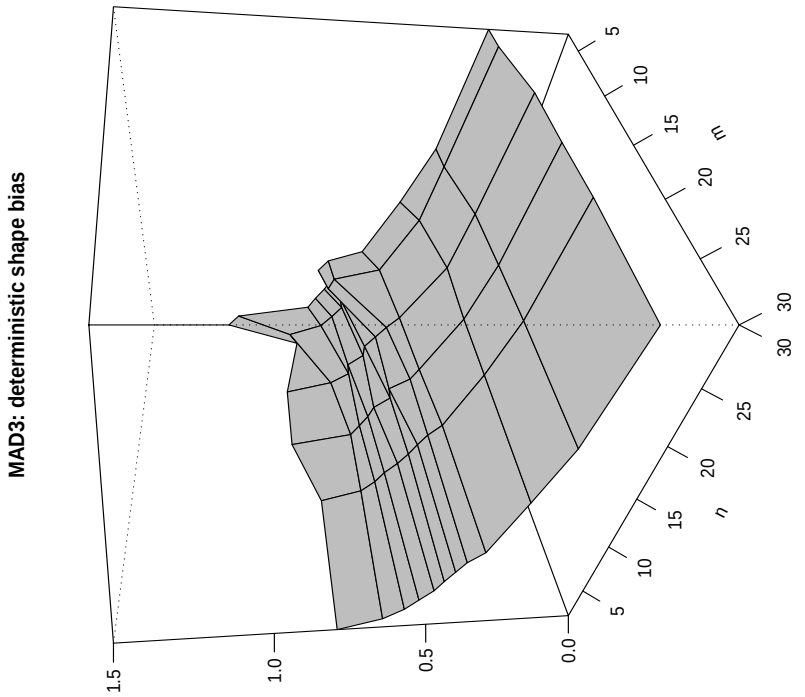


Fig. 10.21. MAD3 error perspective plot similar to Figure 10.17, but with a deterministic perturbation in the shape ($\alpha \to \alpha + 20$) parameter of the Weibull distribution from which the model data is drawn.

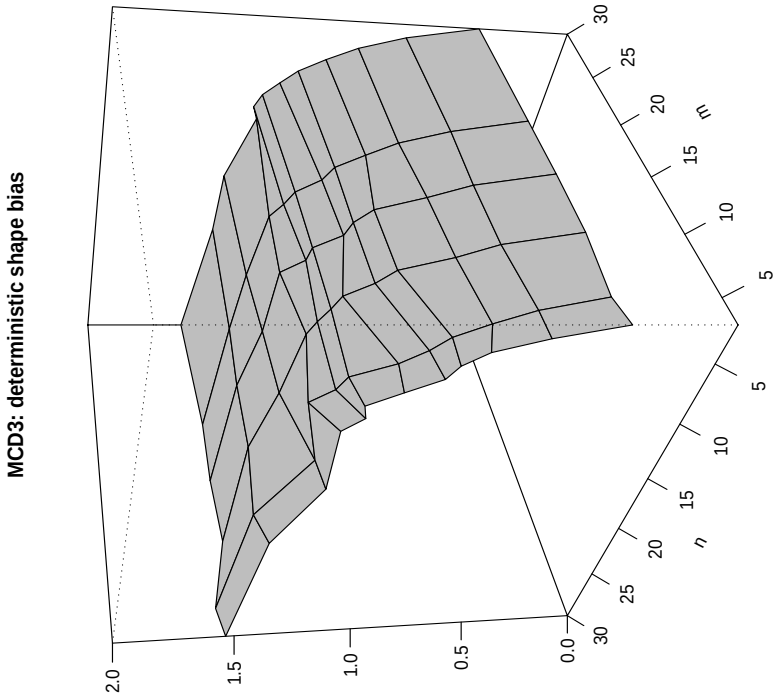


Fig. 10.22. MCD3 error perspective plot similar to Figure 10.18, but with a deterministic perturbation in the shape ($\alpha \rightarrow \alpha + 20$) parameter of the Weibull distribution from which the model data is drawn.

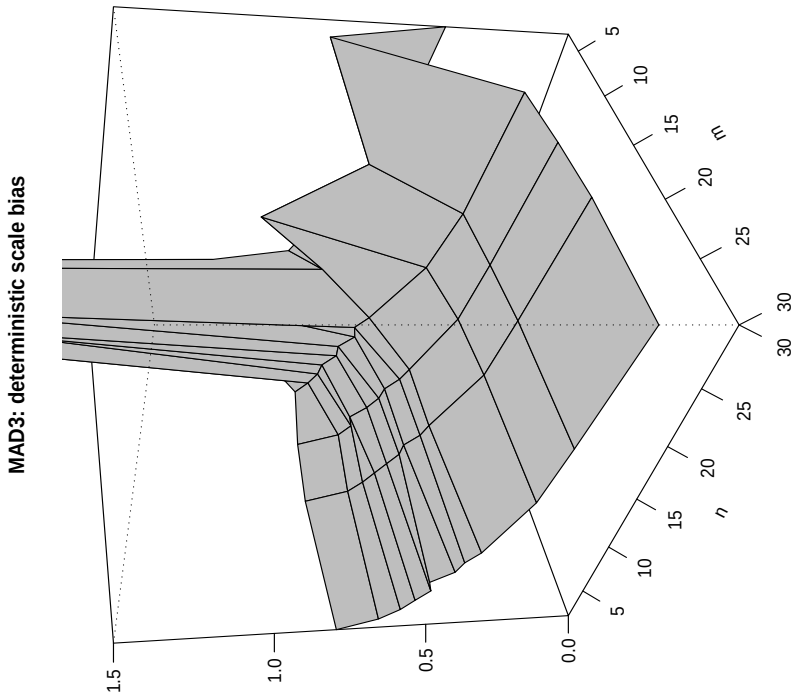


Fig. 10.23. MAD3 error perspective plot similar to Figure 10.17, but with a deterministic perturbation in the scale ($\beta \rightarrow \beta + 20$) parameter of the Weibull distribution from which the model data is drawn.

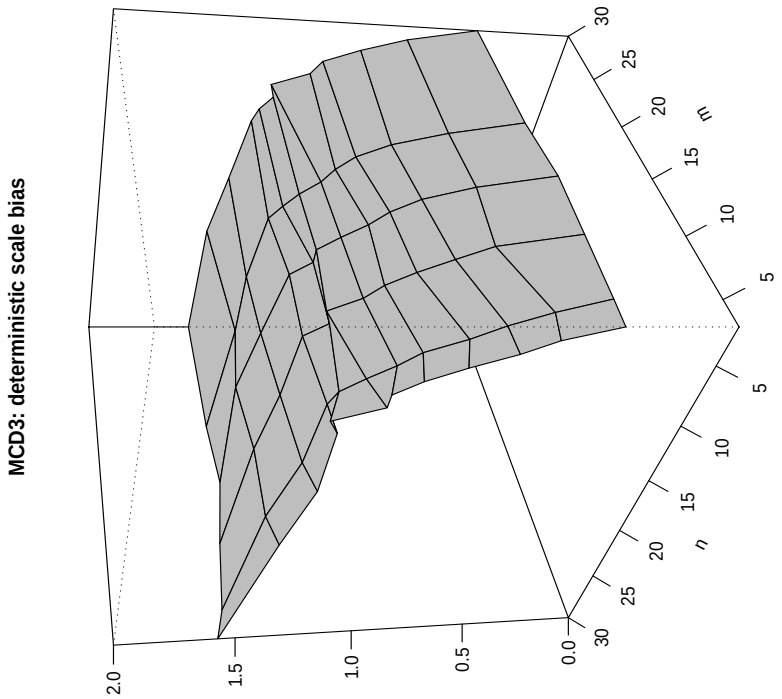


Fig. 10.24. MCD3 error perspective plot similar to Figure 10.18, but with a deterministic perturbation in the scale ($\beta \rightarrow \beta + 20$) parameter of the Weibull distribution from which the model data is drawn.

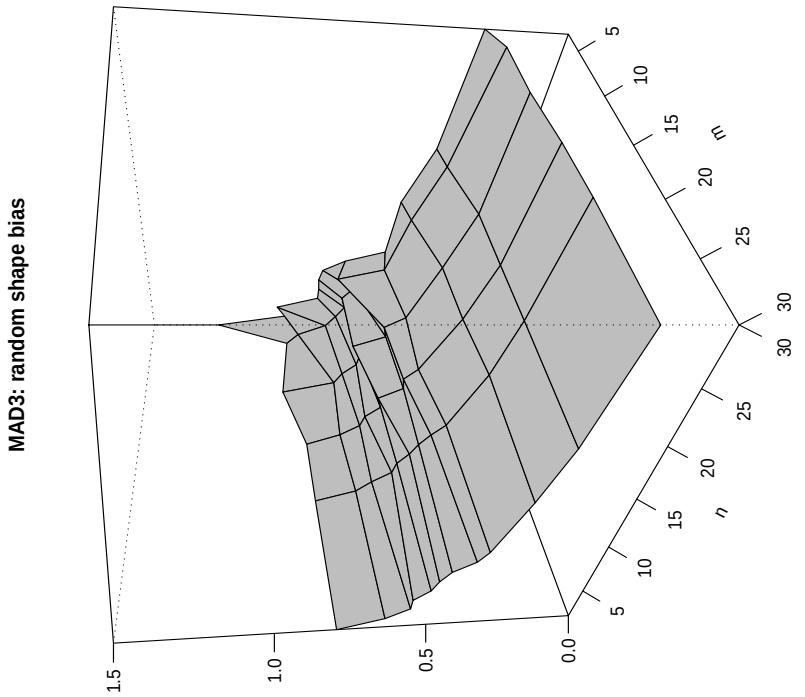


Fig. 10.25. MAD3 error perspective plot similar to Figure 10.17, but with a random perturbation in the shape ($\alpha \to \alpha + 20$) parameter of the Weibull distribution from which the model data is drawn.

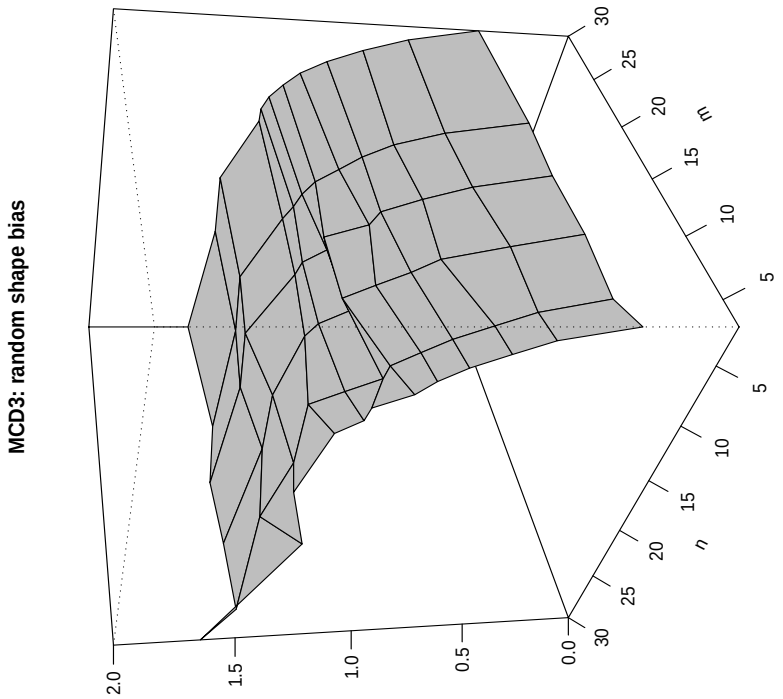


Fig. 10.26. MCD3 error perspective plot similar to Figure 10.18, but with a random perturbation in the shape ($\alpha \rightarrow \alpha + 20$) parameter of the Weibull distribution from which the model data is drawn.

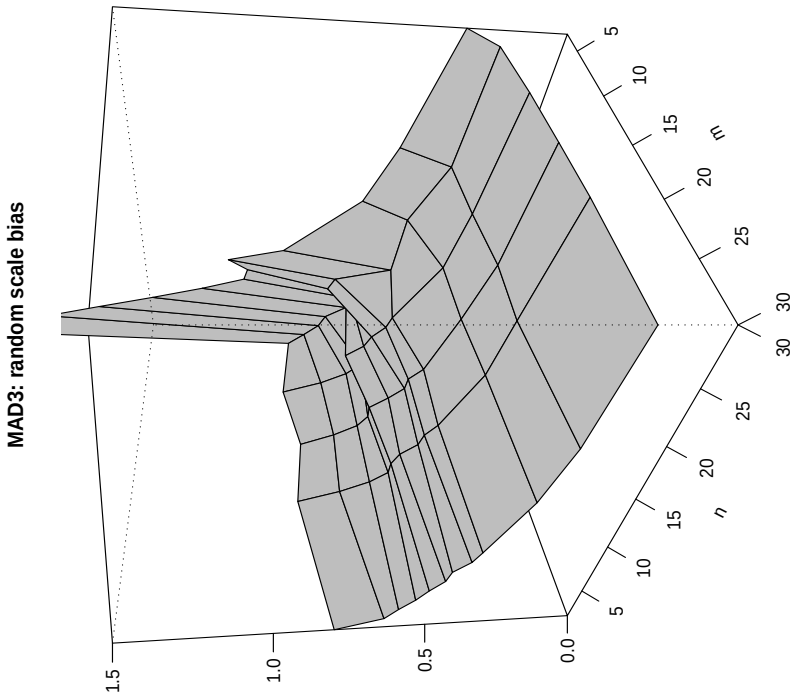


Fig. 10.27. MAD3 error perspective plot similar to Figure 10.17, but with a random perturbation in the scale ($\beta \rightarrow \beta + 20$) parameter of the Weibull distribution from which the model data is drawn.

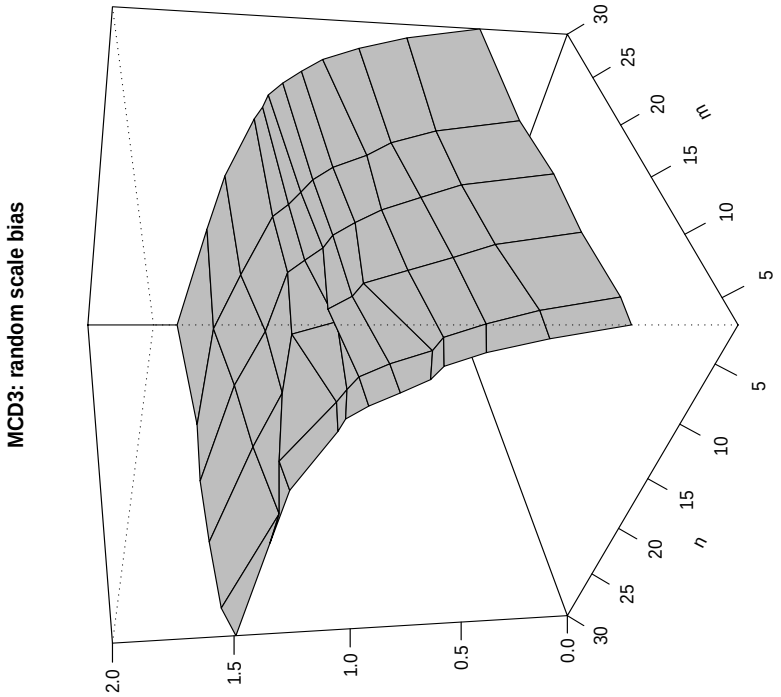


Fig. 10.28. MCD3 error perspective plot similar to Figure 10.18, but with a random perturbation in the scale ($\beta \rightarrow \beta + 20$) parameter of the Weibull distribution from which the model data is drawn.

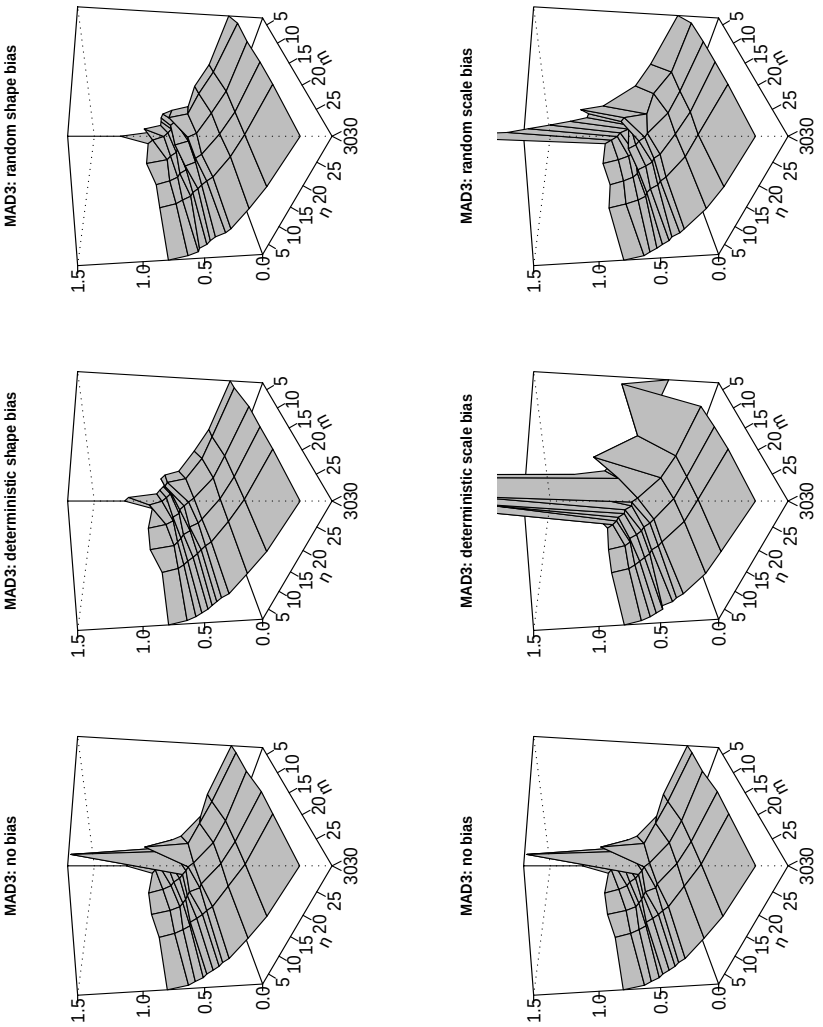


Fig. 10.29. Composite perspective plots of MAD3 error in Figures 10.21-10.28. Figure 10.17 (no bias) is reproduced in column 1 for comparison with deterministic (column 2) and random (column 3) perturbations in the shape and scale parameters of the Weibull distribution from which the model data is drawn.

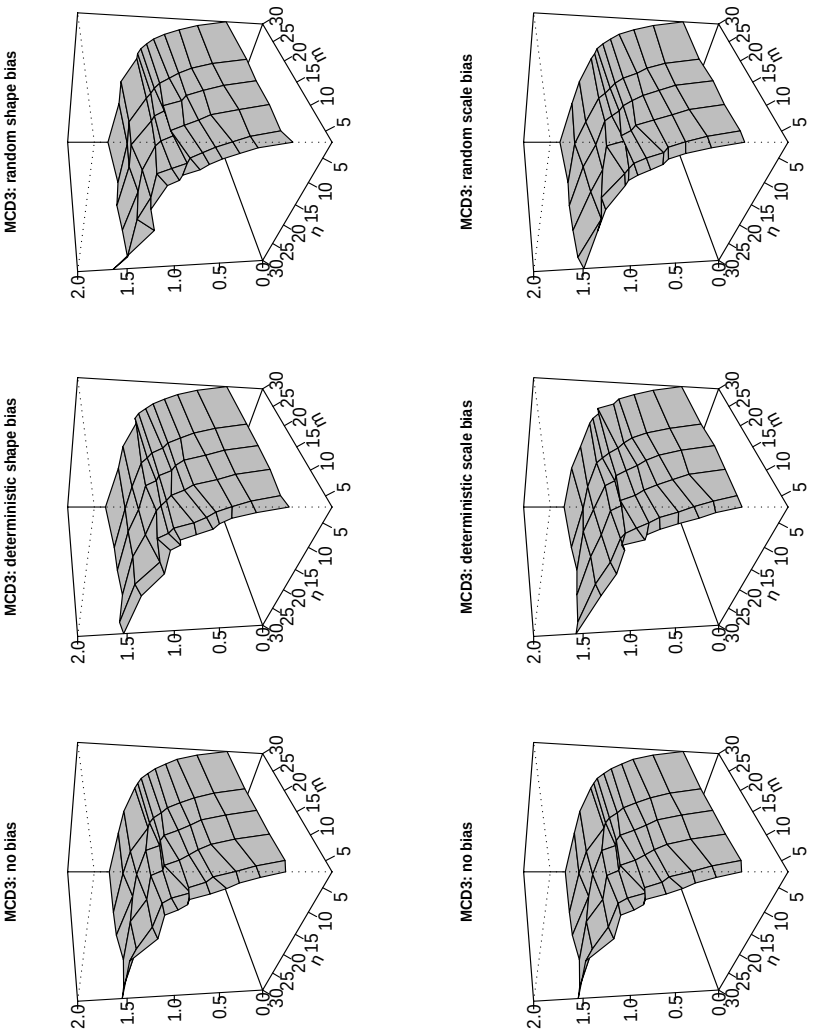


Fig. 10.30. Composite perspective plots of MCD3 error in Figures 10.21-10.28. Figure 10.18 (no bias) is reproduced in column 1 for comparison with deterministic (column 2) and random (column 3) perturbations in the shape and scale parameters of the Weibull distribution from which the model data is drawn.

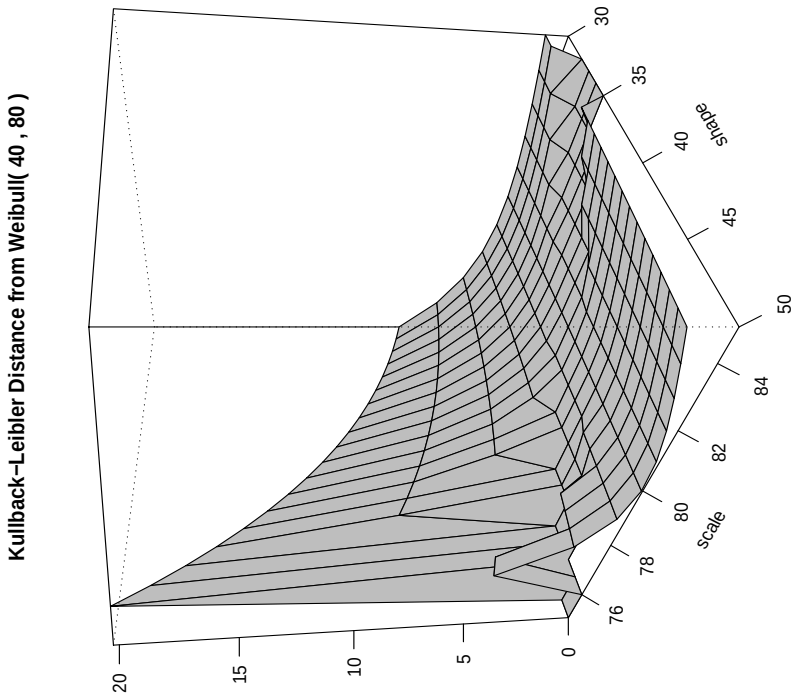


Fig. 10.31. Kullback-Leibler distance of various Weibull distributions from a Weibull(40,80), as a function of the approximating Weibull's shape and scale parameters.

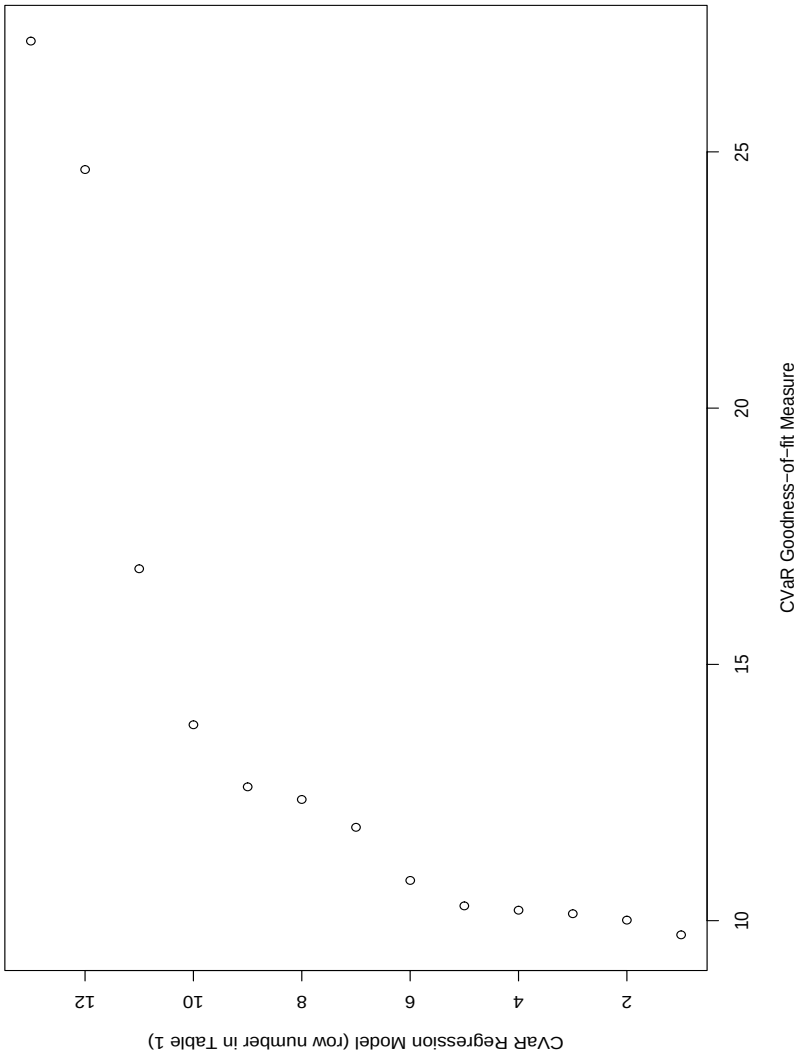


Fig. 10.32. CVaR goodness-of-fit measure plotted against CVaR regression model number (row number in Table 10.1).

Acknowledgements

This effort was jointly accomplished by the Boeing led team and the United States Government under the guidance of NAVAIR. This work was funded by DARPA/DSO and administered by NAVAIR under Technology Investment Agreement N00421-01-3-0098. The program would like to acknowledge the guidance and support of Dr. Leo Christodoulou of DARPA/DSO for this effort.

A CVaR Regression Implementation Details and Flowchart

Mathematical Notation

- Vectors appear in boldface.
- Matrices appear in Roman capitals.
- The superscript $'$ denotes array transposition.
- $\min(a, b)$ and $\max(a, b)$ denote respectively the minimum and maximum of a and b.
- τ denotes any real number between 0 and 1.
- For positive integers N and n with $n \leq N$,

$$\binom{N}{n} = \frac{N!}{n!(N-n)!},$$

  denotes the number of distinct combinations of n items from N.
- z_τ denotes the τth quantile of a Normal distribution with mean 0 and variance 1, i.e.

$$z_\tau = \Phi^{-1}(\tau), \quad \text{where} \quad \Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} \exp\{-x^2/2\} dx.$$

Introduction and Formation of Input Data Arrays

This report provides a detailed explanation of the CVaR regression method proposed in [3] for determination of A-Basis and B-Basis allowables. The details are laid out in a manner that will enable a programmer with only elementary mathematical knowledge to implement the methodology in the form of a computer program. CVaR regression provides a method for integrating different sources of information on failure loads, such as might arise from analytical model and experimental test data. The term **batch** is used here in the broad sense to denote any one single source of variability affecting the test point data. For example: stacking sequence of laminates, type of test, batch number, etc.

Suppose we wish to build a CVaR regression model to integrate information from K analytical models with the physical test data. The available data spans I batches. First define the following:

- Let Y_{ij} be jth failure load value for the ith batch, obtained from test data, $i = 1, \ldots, I$, $j = 1, \ldots, N_i$.
- Let N be the total number of test data points

$$N = \sum_{i=1}^{I} N_i.$$

We then create the data arrays used as the input into the CVaR Regression Module of the next section, as outlined below. The test data can be utilized for covariate and response formation in several ways. We detail two ways, both of which have been used in this report.

Formation of input data arrays: all test data used as both response and covariates

1. Form $\mathbf{y}$, the response vector of length $n = N$ consisting of all failure loads Y_{ij} grouped by batches. The batches and test points within batches, can be arranged in any order, but once decided upon, this order must be maintained throughout.

$$\begin{aligned}\mathbf{y} &= [Y_{11}, Y_{12}, \ldots, Y_{1N_1}, Y_{21}, \ldots, Y_{2N_2}, \ldots, Y_{I1}, \ldots, Y_{IN_I}]' \\ &= [y_1, \ldots, y_n]'.\end{aligned}$$

2. Form, m_i, the sample mean of failure loads for the test data in batch i, for all $i = 1, \ldots, I$.

$$m_i = \frac{1}{N_i} \sum_{j=1}^{N_i} Y_{ij}.$$

3. Form, s_i, the sample standard deviation of failure loads for the test data in batch i, for all $i = 1, \ldots, I$.

$$s_i = \sqrt{\frac{1}{N_i - 1} \sum_{j=1}^{N_i} (Y_{ij} - m_i)^2}.$$

4. Similarly to m_i, form $\mu_i^{(k)}$, the sample mean of failure loads for the data from the kth model in batch i, for all $k = 1, \ldots, K$, and $i = 1, \ldots, I$.
5. Similarly to s_i, form $\sigma_i^{(k)}$, the sample standard deviation of failure loads for the data from the kth model in batch i, for all $k = 1, \ldots, K$, and $i = 1, \ldots, I$.
6. Form the design matrix, X, of CVaR regression. This consists of the test and model sample summary statistics arranged in a matrix of n rows and $l + 1 = 2K + 3$ columns (the first column consists of the number 1). The rows correspond to those in $\mathbf{y}$.

$$\mathrm{X} = \begin{bmatrix} 1 & m_1 & s_1 & \mu_1^{(1)} & \sigma_1^{(1)} & \mu_1^{(2)} & \sigma_1^{(2)} & \ldots & \mu_1^{(K)} & \sigma_1^{(K)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & m_1 & s_1 & \mu_1^{(1)} & \sigma_1^{(1)} & \mu_1^{(2)} & \sigma_1^{(2)} & \ldots & \mu_1^{(K)} & \sigma_1^{(K)} \\ 1 & m_2 & s_2 & \mu_2^{(1)} & \sigma_2^{(1)} & \mu_2^{(2)} & \sigma_2^{(2)} & \ldots & \mu_2^{(K)} & \sigma_2^{(K)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & m_2 & s_2 & \mu_2^{(1)} & \sigma_2^{(1)} & \mu_2^{(2)} & \sigma_2^{(2)} & \ldots & \mu_2^{(K)} & \sigma_2^{(K)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & m_I & s_I & \mu_I^{(1)} & \sigma_I^{(1)} & \mu_I^{(2)} & \sigma_I^{(2)} & \ldots & \mu_I^{(K)} & \sigma_I^{(K)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & m_I & s_I & \mu_I^{(1)} & \sigma_I^{(1)} & \mu_I^{(2)} & \sigma_I^{(2)} & \ldots & \mu_I^{(K)} & \sigma_I^{(K)} \end{bmatrix} \begin{matrix} \\ \\ \uparrow \\ N_1 \text{ rows} \\ \downarrow \\ \uparrow \\ N_2 \text{ rows} \\ \downarrow \\ \\ \vdots \\ \\ \uparrow \\ N_I \text{ rows} \\ \downarrow \end{matrix}$$

Formation of input data arrays: complementary subsets of test data used as response and covariates

In this procedure, the test data points in each batch are partitioned into two disjoint sets. The points in the one set are used as the response, while those in the other are used to form the covariates. This is repeated for all remaining possible ways to partition the test points so as to result in different response and covariate sets. The details are as follows:

1. Repeat the following loop for each batch i, where $i = 1, \ldots, I$:

a) Form all $p_i = \binom{N_i}{w}$ possible partitions of the N_i test points into two disjoint sets consisting of w and $n_i = N_i - w$ points, respectively, where $0 \leq w \leq N_i$, is constant for all i. The points in the first set will be used to form the covariates (the *covariate set*); those in the second will be used as responses (the *response set*).

b) Let $Y^{(q)}_{i(1)}, \ldots, Y^{(q)}_{i(n_i)}$ denote the test points in the qth response set, $q = 1, \ldots, p_i$. Form $\mathbf{y}_i$, the vector of length $n_i p_i$ consisting of the responses ordered as follows,

$$\mathbf{y}_i = [Y^{(1)}_{i(1)}, \ldots, Y^{(1)}_{i(n_i)}, Y^{(2)}_{i(1)}, \ldots, Y^{(2)}_{i(n_i)}, \ldots, Y^{(p_i)}_{i(1)}, \ldots, Y^{(p_i)}_{i(n_i)}]'.$$

c) Let $Y^{(q)}_{i(n_i+1)}, \ldots, Y^{(q)}_{i(N_i)}$ denote the test points in the qth covariate set, $q = 1, \ldots, p_i$. Form $m^{(q)}_i$ and $s^{(q)}_i$, the sample mean and standard deviation of the test data in the qth covariate set,

$$m^{(q)}_i = \frac{1}{w} \sum_{j=n_i+1}^{N_i} Y^{(q)}_{i(j)},$$

$$s^{(q)}_i = \sqrt{\frac{1}{w-1} \sum_{j=n_i+1}^{N_i} \left(Y^{(q)}_{i(j)} - m^{(q)}_i \right)^2}.$$

If $w = 0$, the test data is not used as covariates (only the model). If $w = 1$, then $s^{(q)}_i = 0$. If $w = N_i$, we use all the test data as both response and covariates as in the preceding subsection.

d) Form $\mu^{(k)}_i$ and $\sigma^{(k)}_i$, the sample mean and standard deviation of failure loads for the data from the kth model in batch i, for all $k = 1, \ldots, K$.

e) Form the design matrix, X_i, for batch i. This consists of the test and model sample summary statistics arranged in a matrix of $n_i p_i$ rows and $2K + 3$ columns (the first column consists of the number 1). The rows correspond to those in $\mathbf{y}_i$.

$$\mathrm{X}_i = \begin{bmatrix} 1 & m^{(1)}_i & s^{(1)}_i & \mu^{(1)}_i & \sigma^{(1)}_i & \mu^{(2)}_i & \sigma^{(2)}_i & \ldots & \mu^{(K)}_i & \sigma^{(K)}_i \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & m^{(1)}_i & s^{(1)}_i & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & m^{(2)}_i & s^{(2)}_i & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & m^{(2)}_i & s^{(2)}_i & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & m^{(p_i)}_i & s^{(p_i)}_i & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & m^{(p_i)}_i & s^{(p_i)}_i & \mu^{(1)}_i & \sigma^{(1)}_i & \mu^{(2)}_i & \sigma^{(2)}_i & \ldots & \mu^{(K)}_i & \sigma^{(K)}_i \end{bmatrix} \begin{matrix} \\ \\ \uparrow \\ n_i \text{ rows} \\ \downarrow \\ \uparrow \\ n_i \text{ rows} \\ \downarrow \\ \\ \vdots \\ \\ \uparrow \\ n_i \text{ rows} \\ \downarrow \\ \\ \end{matrix}$$

2. Form $\mathbf{y}$, the response vector of length $n = \sum_{i=1}^{I} n_i p_i$ comprised of the vertical concatenation of the batch response vectors,

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_I \end{bmatrix}.$$

3. Form the design matrix, X, of CVaR regression. This is a matrix of n rows and $l + 1 = 2K + 3$ columns, comprised of the vertical concatenation of the batch design matrices X_i.

$$\mathrm{X} = \begin{bmatrix} \mathrm{X}_1 \\ \vdots \\ \mathrm{X}_I \end{bmatrix}.$$

Example

Suppose batch i has $\{10.1, 12.3, 14.5, 16.7\}$ as test data, and $\{12.4, 16.8\}$ as model data. We will illustrate the calculations in item 1(a). Here $K = 1$, $N_i = 4$, and suppose we choose $w = 2$. Then $p_i = \binom{4}{2} = 6$, and X_i will be of dimension (12×5). We obtain,

$$\mu_i^{(1)} = (12.4 + 16.8)/2 = 14.6, \qquad \text{and} \qquad \sigma_i^{(1)} = 16.8 - 12.4 = 4.4.$$

For the first covariate set, $\{14.5, 16.7\}$, we obtain

$$m_i^{(1)} = (14.5 + 16.7)/2 = 15.6, \qquad \text{and} \qquad s_i^{(1)} = 16.7 - 14.5 = 2.2.$$

Computing $m_i^{(2)}, \ldots, m_i^{(6)}$ and the corresponding $s_i^{(2)}, \ldots, s_i^{(6)}$ similarly, gives eventually,

$$\mathbf{y}_i = \begin{bmatrix} 10.1 \\ 12.3 \\ 10.1 \\ 14.5 \\ 10.1 \\ 16.7 \\ 12.3 \\ 14.5 \\ 12.3 \\ 16.7 \\ 14.5 \\ 16.7 \end{bmatrix}, \qquad \mathrm{X}_i = \begin{bmatrix} 1 & 15.6 & 2.2 & 14.6 & 4.4 \\ 1 & 15.6 & 2.2 & 14.6 & 4.4 \\ 1 & 14.5 & 4.4 & 14.6 & 4.4 \\ 1 & 14.5 & 4.4 & 14.6 & 4.4 \\ 1 & 13.4 & 2.2 & 14.6 & 4.4 \\ 1 & 13.4 & 2.2 & 14.6 & 4.4 \\ 1 & 13.4 & 6.5 & 14.6 & 4.4 \\ 1 & 13.4 & 6.5 & 14.6 & 4.4 \\ 1 & 12.3 & 4.4 & 14.6 & 4.4 \\ 1 & 12.3 & 4.4 & 14.6 & 4.4 \\ 1 & 11.2 & 2.2 & 14.6 & 4.4 \\ 1 & 11.2 & 2.2 & 14.6 & 4.4 \end{bmatrix}.$$

The CVaR Regression Module

This module takes on the following inputs from the previous section:

Inputs:

- $\mathbf{y}$: the test data (vector of length n);
- X: the design matrix (n by $l + 1$);
- τ: the probability level of the desired quantile ($0 \le \tau \le 0.5$).

Here for simplicity we describe only the case when $0 \le \tau \le 0.5$. The $\tau > 0.5$ case can be treated symmetrically; see Appendix C in [3]. From these inputs, the following are computed:

1. Let $\mathbf{x}_r$ designate the vector of length $l + 1$ consisting of the rth row of X. Defining $\overline{\mathrm{X}}$ to be X without the first column, Let $\bar{\mathbf{x}}_r$ designate the vector of length l consisting of the rth row of $\overline{\mathrm{X}}$. With this notation, we can write

$$\mathrm{X} = \begin{bmatrix} \mathbf{x}_1' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}, \qquad \text{and} \qquad \overline{\mathrm{X}} = \begin{bmatrix} \bar{\mathbf{x}}_1' \\ \vdots \\ \bar{\mathbf{x}}_n' \end{bmatrix}.$$

2. Let $\mathbf{c}$ denote the vector of length $l + 1$ of CVaR regression coefficients

$$\mathbf{c} = [c_0, c_1, \ldots, c_l]'.$$

In addition, let $\bar{\mathbf{c}}$ be $\mathbf{c}$ without c_0. Let $\mathcal{P}_\tau^2(\bar{\mathbf{c}})$ designate the goodness-of-fit criterion (or objective function) for the fitted τth quantile CVaR regression surface, evaluated at $\bar{\mathbf{c}}$ defined as,

$$\mathcal{P}_\tau^2(\overline{\mathbf{c}}) = \widehat{\mathrm{CVaR}}_{1-\tau}^{\Delta}(-\overline{\varepsilon}) = \frac{1-\tau}{\tau}\widehat{\mathrm{CVaR}}_{\tau}^{\Delta}(\overline{\varepsilon}).$$

(See formulas (22) and (23) in [3].) $\widehat{\mathrm{CVaR}}_{1-\tau}^{\Delta}(-\overline{\varepsilon})$ is the average of the largest $n\tau$ elements in the list of centered negative residuals,

$$-\overline{\varepsilon}_i = -\left[y_i - \overline{\mathbf{x}}_i'\overline{\mathbf{c}} - \frac{1}{n}\sum_{j=1}^{n}(y_j - \overline{\mathbf{x}}_j'\overline{\mathbf{c}})\right], \quad i = 1, \ldots, n.$$

As such, $\widehat{\mathrm{CVaR}}_{1-\tau}^{\Delta}(-\overline{\varepsilon})$ is an estimate of the distance between the mean and the mean of the right tail beyond the $(1-\tau)$th quantile of the CVaR regression residual distribution. If $n\tau$ is not an integer, we must count a fraction of an element in the list of residuals. For example, suppose that $n = 5$, $\tau = 0.3$, and the list of centered negative residuals is $\{6, 4, 12, 1, 3\}$. In this case, $\widehat{\mathrm{CVaR}}_{1-\tau}^{\Delta}(-\overline{\varepsilon})$ equals the average of the largest $5(0.3) = 1.5$ elements. Since the two largest elements are 6 and 12, we have

$$\widehat{\mathrm{CVaR}}_{1-\tau}^{\Delta}(-\overline{\varepsilon}) = \frac{1(12) + 0.5(6)}{1.5} = \frac{15}{1.5} = 10.$$

3. Let $\hat{\overline{\mathbf{c}}}(\tau)$ denote the minimizer of $\mathcal{P}_\tau^2(\overline{\mathbf{c}})$ over all $\overline{\mathbf{c}} \in \mathscr{R}^l$, at quantile τ, i.e.

$$\hat{\overline{\mathbf{c}}}(\tau) = \arg\min_{\overline{\mathbf{c}} \in \mathscr{R}^l} \mathcal{P}_\tau^2(\overline{\mathbf{c}}). \tag{10.4}$$

   For details on how this minimization can be performed efficiently, see Section A.
4. Define

$$\hat{c}_0(\tau) = \widehat{\mathrm{VaR}}_{1-\tau}\left(-\hat{\overline{\varepsilon}}\right) - \frac{1}{n}\sum_{j=1}^{n}(y_j - \overline{\mathbf{x}}_j'\hat{\overline{\mathbf{c}}}(\tau)).$$

   If $\{-\overline{\varepsilon}_{(1)} \le -\overline{\varepsilon}_{(2)} \le \cdots \le -\overline{\varepsilon}_{(n)}\}$ denote the ordered centered negative residuals, we define

$$\widehat{\mathrm{VaR}}_{1-\tau}(-\overline{\varepsilon}) = -\overline{\varepsilon}_{(r)},$$

   where r is the unique integer satisfying $r - 1 \le n(1-\tau) < r$. In the above example with $n = 5$ and $\tau = 0.3$, $(1-\tau)n = 3.5$, and therefore $\widehat{\mathrm{VaR}}_{1-\tau}(-\overline{\varepsilon})$ is the 4th ordered element, i.e. 6.
5. Denote by $\hat{\mathbf{c}}(\tau) = [\hat{c}_0(\tau), \hat{\overline{\mathbf{c}}}(\tau)']'$, the estimator of $\mathbf{c}$ in CVaR regression fitted at quantile level τ.
6. Let $\hat{\varepsilon}_i^*$ denote the ith residual from the CVaR regression fit, i.e.

$$\hat{\varepsilon}_i^* = y_i - \mathbf{x}_i'\hat{\mathbf{c}}(\tau),$$

   and let $\hat{\varepsilon}_{(1)}^* \le \hat{\varepsilon}_{(2)}^* \le \cdots \le \hat{\varepsilon}_{(n)}^*$ denote these n ordered residuals. Define

$$\hat{F}_n^{-1}(\tau) = \widehat{\mathrm{VaR}}_{\tau}(\hat{\varepsilon}^*) = \hat{\varepsilon}_{(r)}^*,$$

   where r is the unique integer satisfying $r - 1 \le n\tau < r$.
7. Let $h_n(\tau)$ be the bandwidth at quantile τ defined by,

$$h_n(\tau) = \left[\frac{3z_{0.975}^2 \exp\{-z_\tau^2\}}{4n\pi(2z_\tau^2 + 1)}\right]^{1/3}.$$

8. Let $\hat{s}_n(\tau)$ be the sparsity function at quantile τ defined by,

$$\hat{s}_n(\tau) = \frac{\hat{F}_n^{-1}(\tau + h_n(\tau)) - \hat{F}_n^{-1}(\tau - h_n(\tau))}{2h_n(\tau)}.$$

9. Let $\hat{\Omega}_n(\tau)$ be the asymptotic covariance matrix at quantile τ defined by,

$$\hat{\Omega}_n(\tau) = \tau(1-\tau)(\mathrm{X}'\mathrm{X})^{-1}\hat{s}_n(\tau)^2.$$

   It has $l + 1$ columns and $l + 1$ rows.

The following is the desired output from the CVaR Regression Module:

Outputs:

- $\hat{\mathbf{c}}(\tau)$: the optimal coefficients;
- $\mathcal{P}_\tau^2\left(\hat{\bar{\mathbf{c}}}(\tau)\right)$: the CVaR goodness-of-fit measure;
- $\hat{\Omega}_n(\tau)$: the asymptotic covariance matrix.

Final Quantities of Interest

Using the outputs from the CVaR Regression Module, we can calculate several quantities. Suppose that the summary statistics for a particular batch of interest are available. We can arrange them into a vector as follows,

$$\mathbf{x} = [1, m_1, s_1, \mu_1^{(1)}, \sigma_1^{(1)}, \mu_1^{(2)}, \sigma_1^{(2)}, \ldots, \mu_1^{(K)}, \sigma_1^{(K)}]'.$$

From this we can compute the following.

τth quantile of failure load at $\mathbf{x}$. The estimated value for the τth quantile of the failure load when the batch values are as in $\mathbf{x}$, is given by

$$\hat{Q}_Y(\tau) = \hat{c}_0(\tau) + \hat{c}_1(\tau)m_1 + \hat{c}_2(\tau)s_1 + \hat{c}_3(\tau)\mu_1^{(1)} + \hat{c}_4(\tau)\sigma_1^{(1)} + \cdots + \hat{c}_{2K+1}(\tau)\mu_1^{(K)} + \hat{c}_{2K+2}(\tau)\sigma_1^{(K)}. \quad (10.5)$$

For example, the estimated 10th quantile of the failure load is $\hat{Q}_Y(0.10)$.

Goodness-of-fit of τ-CVaR regression. The overall goodness-of-fit measure of the τth quantile CVaR regression factor model fitted to the data, is simply $\mathcal{P}_\tau^2(\hat{\mathbf{c}}(\tau))$. For example, the goodness-of-fit of a 10th quantile CVaR regression fit is $\mathcal{P}_\tau^2(\hat{\mathbf{c}}(0.10))$.

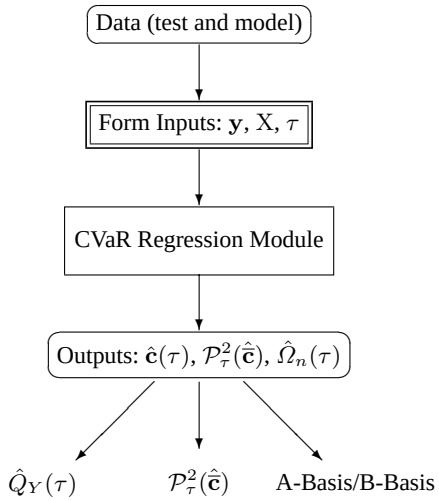
A-Basis. The A-Basis estimate when the batch values are as in $\mathbf{x}$, is given by

$$\mathbf{x}'\hat{\mathbf{c}}(0.01) + z_{0.05}\sqrt{\mathbf{x}'\hat{\Omega}_n(0.01)\mathbf{x}}. \quad (10.6)$$

B-Basis. The B-Basis estimate when the batch values are as in $\mathbf{x}$, is given by

$$\mathbf{x}'\hat{\mathbf{c}}(0.10) + z_{0.05}\sqrt{\mathbf{x}'\hat{\Omega}_n(0.10)\mathbf{x}}. \quad (10.7)$$

CVaR Regression Flowchart Diagram



Minimization of CVaR: Reduction to Linear Programming

The minimization problem

$$\min_{\overline{\mathbf{c}}\in\mathscr{R}^l} \mathcal{P}^2_\tau(\overline{\mathbf{c}}) = \min_{\overline{\mathbf{c}}\in\mathscr{R}^l} \widehat{\mathrm{CVaR}}^{\Delta}_{1-\tau}\left(-\overline{\varepsilon}\right), \tag{10.8}$$

can be reduced to the following linear programming problem; see [2]:

$$\min\left\{\eta + \frac{1}{n\tau}\sum_{i=1}^{n} z_i\right\},$$

where

$$z_i \geq -y_i + \frac{1}{n}\sum_{i=1}^{n} y_i + \left(\bar{\mathbf{x}}_i' - \frac{1}{n}\sum_{i=1}^{n}\bar{\mathbf{x}}_i'\right)\overline{\mathbf{c}} - \eta, \quad i = 1,\ldots,n,$$
$$z_i \geq 0, \quad i = 1,\ldots,n,$$

and

$$\eta \in \mathscr{R}, \quad \overline{\mathbf{c}} \in \mathscr{R}^l.$$

The minimization is conducted with respect to $\overline{\mathbf{c}} = [\overline{c}_1,\ldots,\overline{c}_l]' \in \mathscr{R}^l$; $\eta \in \mathscr{R}$; and $z_i \in \mathscr{R}$, $i = 1,\ldots,n$.

Let $\hat{\overline{\mathbf{c}}} = [\hat{\overline{c}}_1,\ldots,\hat{\overline{c}}_l]'$, $\hat{\eta}$, and $\hat{z}_1,\ldots,\hat{z}_n$ denote the minimizers. The minimization simultaneously minimizes $\widehat{\mathrm{CVaR}}^{\Delta}_{1-\tau}\left(-\overline{\varepsilon}\right)$ and finds $\widehat{\mathrm{VaR}}_{1-\tau}\left(-\overline{\varepsilon}\right)$ at the optimal $-\overline{\varepsilon}$. In other words,

$$\min_{\overline{\mathbf{c}}\in\mathscr{R}^l} \widehat{\mathrm{CVaR}}^{\Delta}_{1-\tau}\left(-\overline{\varepsilon}\right) = \widehat{\mathrm{CVaR}}^{\Delta}_{1-\tau}\left(-\hat{\overline{\varepsilon}}\right) = \mathcal{P}^2_\tau(\hat{\overline{\mathbf{c}}}) = \hat{\eta} + \frac{1}{n\tau}\sum_{i=1}^{n}\hat{z}_i,$$

and

$$\widehat{\mathrm{VaR}}_{1-\tau}\left(-\hat{\overline{\varepsilon}}\right) = \hat{\eta}.$$

References

[1] Kullback, S., and Leibler, R.A. (1951), "On information and sufficiency", *Annals of Mathematical Statistics* 22, 79-86.

[2] Rockafellar R.T. and S. Uryasev. (2002), "Conditional Value-at-Risk for General Loss Distributions", *Journal of Banking and Finance*, 26-7, 1443-1471.

[3] Uryasev S., Trindade A., and Uryasev Jr., S. (2003), "Combining Model and Test Data for Optimal Determination of Percentiles and Allowables: CVaR Regression Approach, Part I". American Optimal Decisions, Inc., report conducted in the framework of AIM-C program.