Chapter 9

Extension to Multiple Regression

9.A Introduction

The multiple regression extension of the Outlier Resistant Model Robust Regression (ORMRR) methodology described in Chapter 6 is relatively straightforward with few considerations necessary. Little attention has been given to the performance of nonparametric regression estimators in relatively small sample sizes when multiple regressor variables exist. This chapter is devoted to one approach of extending nonparametric regression methodology to higher dimensions in the context of our ORMRR modelling procedure. It is reasonable to assume that this methodology (that performs well in many applied situations) must be developed for the nonparametric or semiparametric procedures to gain popularity in application.

Intuitively, the difficulty in the extension of robust nonparametric procedures to the multiple regressor space stems from the fact that there is no global underlying model that is specified. Thus the likelihood of a nonparametric regression technique identifying the correct trend relies heavily on the distribution and number of observations available in the regressor space. This difficulty is compounded by the fact that residual based weighting is incorporated, making it difficult for the technique to discriminate between good and bad data without the benefit of several data points in a close neighborhood of the point in question.

9.B Extension of Methodology to Multiple Regression Technique

The extension of M-Regression to higher dimensions is well documented. It is simply the augmentation of the model matrix $\mathbf{X}^P$ with additional columns that correspond to model terms. The residual based weighting scheme utilized in the M-Regression fit remains the same, with observations receiving weights based on the size of their corresponding residuals. For a more

detailed discussion on M-Regression for multiple regressor variables, see, for example, Huber (1981), Myers (1990), or Staudte and Sheather (1990).

Since the robust nonparametric portion of the ORMRR estimator is analogous to a local M-Estimation fit, it is intuitive that the extension of this component of ORMRR be similar to that of M-Regression. The element that is unique is the *local* nature of the fit. From Chapter 4, the *neighborhood* weight for the j^{th} observation when predicting at location i can be expressed in the form

$$h_{ij}^{KER} = \frac{n^{-1}b_n^{-1}K\left(\frac{d_{ij}}{b_n}\right)}{n^{-1}b_n^{-1}\sum_{k=1}^{n}K\left(\frac{d_{ik}}{b_n}\right)},$$

where d_{ij} is some norm or distance measure between x_i and x_j (specifically, $d_{ij} = x_i - x_j$ in the single regressor case).

In the multiple regressor case, the parametric vector at the i^{th} location is denoted by $\mathbf{x}_i^P$, which takes, for example, the form $\mathbf{x}_i^{P\prime} = (1\ x_{1i}\ \ x_{2i}\ \ x_{1i}\cdot x_{2i}\ \ x_{1i}^2\ \ x_{2i}^2)$ if the user's specified model is a full quadratic in two regressors. This is the vector that is used to determine the parametric fitted value to the raw data: $\hat{y}_i^M = \mathbf{x}_i^{P\prime}\hat{\boldsymbol{\beta}}^M$. In addition, the nonparametric vector at the i^{th} location, $\mathbf{x}_i^{NP}$, requires modification to include the multiple regressors. For example, if fitting a local linear regression model (or robust variation thereof) in two regressors, this vector takes the form $\mathbf{x}_i^{NP\prime} = (1\ x_{1i}\ \ x_{2i})$ which is used to calculate $\hat{y}_i^{LLR} = \mathbf{h}_i^{LLR\prime}\mathbf{y}$ where

$$\mathbf{h}_i^{LLR\prime} = x_i^{NP\prime}(\mathbf{X}^{NP\prime} < h_{ij}^{KER} > \mathbf{X}^{NP})^{-1}\mathbf{X}^{NP\prime} < h_{ij}^{KER} > .$$

The neighborhood weights for the nonparametric portion of the multiple regression fit will be the same as above, using the Euclidean distance between the vectors $\mathbf{x}_i^*$ and $\mathbf{x}_j^*$, where, for example, $\mathbf{x}_i^* = (x_{1i}\ \ x_{2i})$ for two regressor variables. The Euclidean distance between these two vectors is given by

$$d_{ij} = \sqrt{(\mathbf{x}_i^* - \mathbf{x}_j^*)'(\mathbf{x}_i^* - \mathbf{x}_j^*)}\,.$$

Thus, since the weight function used in this research is the simplified normal, the kernel weights are of the form

$$h_{ij}^{KER} = \frac{e^{\frac{-d_{ij}^2}{b_n}}}{\sum_{k=1}^{n} e^{\frac{-d_{ij}^2}{b_n}}}\,.$$

The assumption is also made, as in the single regressor case, that there are no outliers in the regressor space (high leverage points), making the use of Euclidean distance as the norm a reasonable selection.

Using these definitions for the components of the ORMRR method, the model then becomes

$$\begin{aligned} y_i &= m(\mathbf{x}_i) + \boldsymbol{\varepsilon}_i \\ &= \mathbf{x}_i^{P\prime}\boldsymbol{\beta} + f(\mathbf{x}_i) + \boldsymbol{\varepsilon}_i \end{aligned}\,.$$

The following section offers one specific example of how the estimator performs for a given set of data and is in no way meant to serve as a generalization of the performance of the estimator in other situations. A comparison is offered, however, with M-Regression and Loess in order to give some insight into how these estimators differ in the multivariate setting.

9.C Example

For the sake of simplicity, an example with two regressor variables is presented. Below is the true underlying model from which the data were generated:

$$E(y|\mathbf{x}) = 2x_1^2 - 2x_2^2 + 4\gamma \cdot \sin(.75 x_1 x_2)\,. \qquad (9.1)$$

Note that the function is a true quadratic that can be distorted by a sine function. As before, the value of γ is contained in the interval [0,1] and determines the degree of departure of the true underlying from a purely quadratic model in the regressors. Figure 9.1 is a plot of the true response function for $\gamma = 0$ (a true quadratic surface).

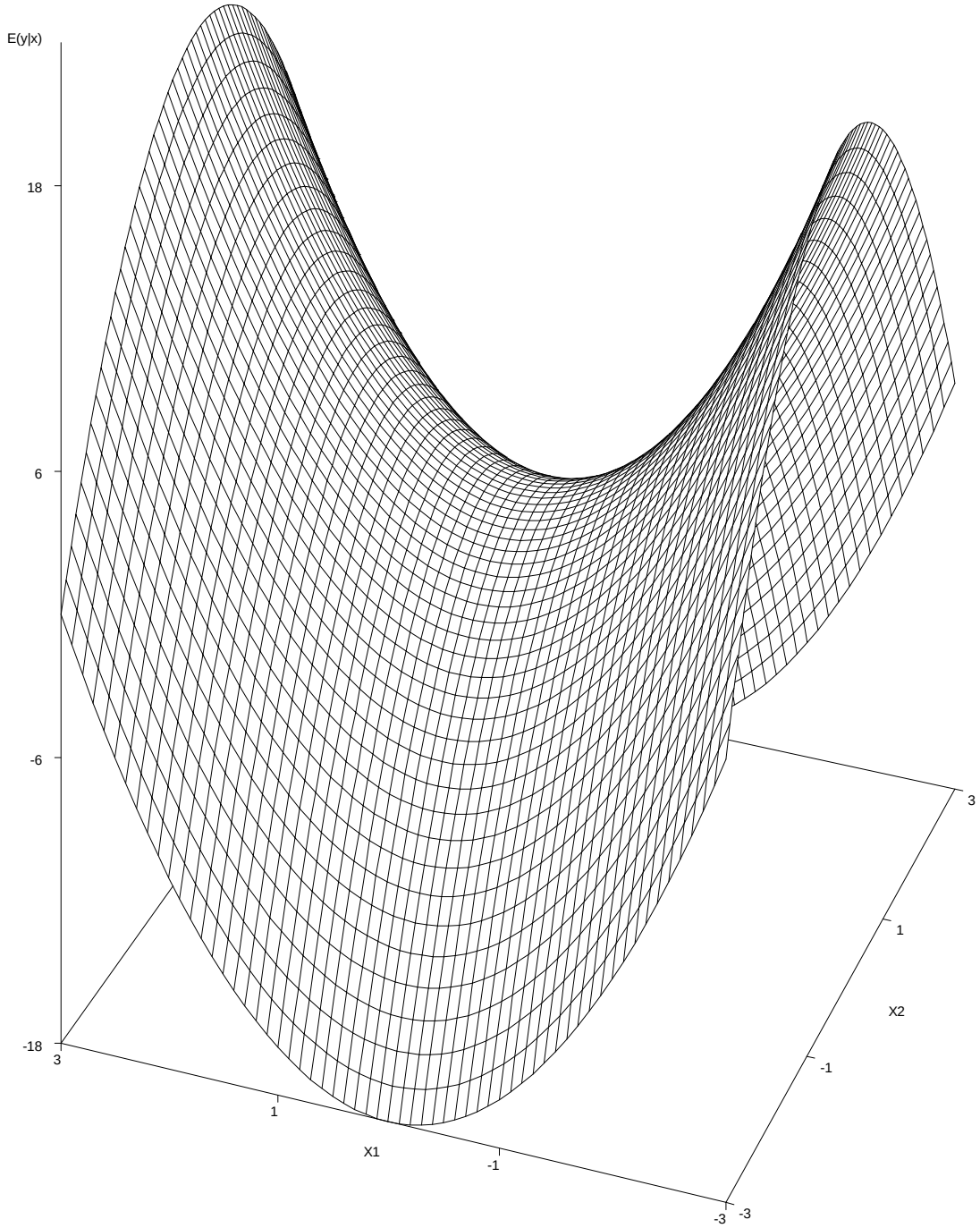

Figure 9.1 Plot of model $E(y|\mathbf{x}) = 2x_1^2 - 2x_2^2 + 4\gamma \cdot \sin(.75x_1 x_2)$ with $\gamma = 0$ (no model misspecification for quadratic user's model).

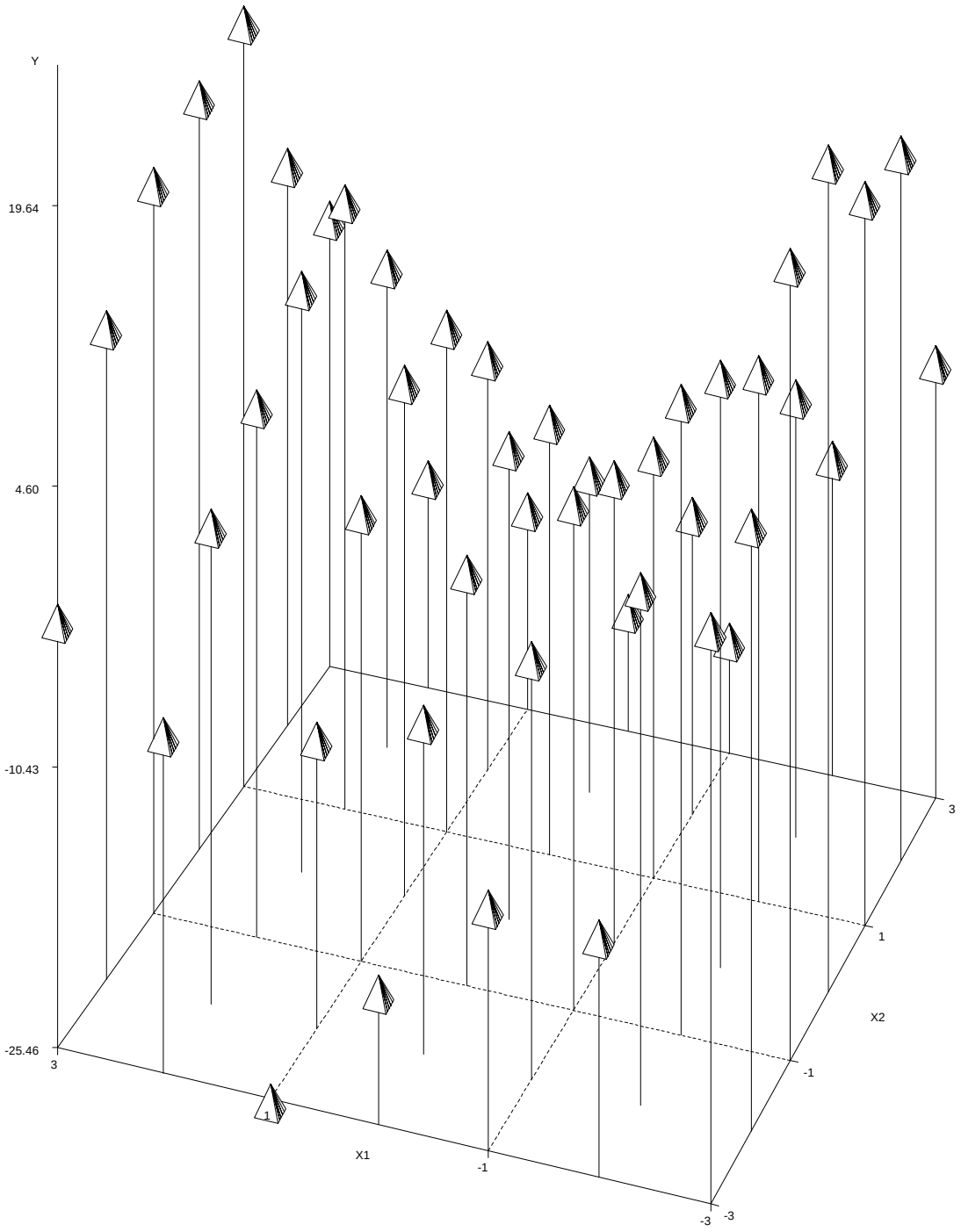

Figure 9.2 Plot of data generated from model $E(y|\mathbf{x}) = 2x_1^2 - 2x_2^2 + 4\gamma \cdot \sin(.75x_1 x_2)$ with $\gamma = 0.75$ and error distribution CN(0.10, 1.0, 5.0).

Since the situation of interest is when the model is misspecified, a set of data generated from a model that has a value of γ that is different from 0 will be used in order to compare the techniques in a multivariate setting. Consider the model above with $\gamma = 0.75$, corresponding to a moderately large amount of model misspecification if the user attempts to fit a full quadratic model.

The full quadratic (linear terms, quadratic terms, and cross-product terms) will be employed as the user's specified model. In response surface methodology, it is a common assumption that the full quadratic model is reasonably sufficient for most surfaces. Based on this assumption, and the plot of the data in Figure 9.2, it appears reasonable to assume that many users would believe a full quadratic model to be sufficient.

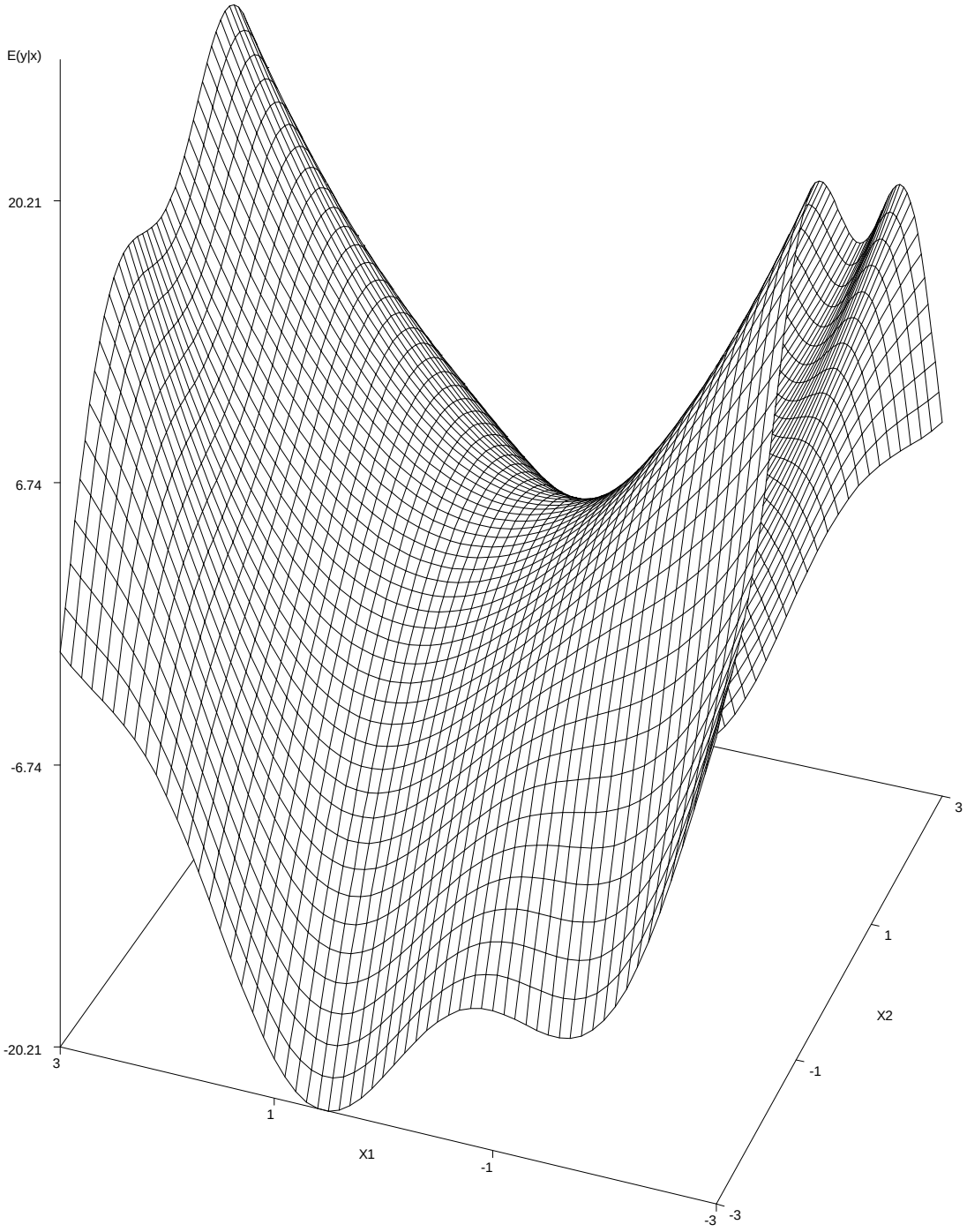

Figure 9.3 Plot of model $E(y|\mathbf{x}) = 2x_1^2 - 2x_2^2 + 4\gamma \cdot \sin(.75x_1 x_2)$
with $\gamma = 0.75$ (moderate amount of model misspecification for quadratic user's model).

Consider an analysis of the fit to the multiple regressor data set generated from the model described in (9.1) and graphed in Figure 9.2. The error distribution that was used to generate the data set was CN(0.1, 1.0, 5.0). Equally spaced observations were also used, based on a 7x7 grid in the two regressor variables (49 data points). Predicted values were obtained for a 41x41 grid (1681 locations) in order to obtain a detailed plot of the fitted surface, and to calculate the emse of the competing procedures for this data set (recall from Chapter 7 that emse was the estimated mean squared error, an average of the squared distances of the predicted value from the true mean value at a grid of locations in the regressor space).

The parameters chosen for the ORMRR fit were not theoretical nor data driven, but were chosen as those used in the single regressor case for n = 50. This was a bandwidth of $b_n = 0.10$ and a mixing parameter of $\lambda = 1.0$.

Figure 9.4 is a plot of the estimated response curve generated by the ORMRR procedure, which closely resembles the true surface in Figure 9.3. A plot of the smoothed residuals (obtained using RLLR) is given in Figure 9.5, and resembles the sine curve that serves as the model misspecification portion of the example being considered.

Figures 9.6 and 9.7 are plots of the fitted surface using Loess with parameter values f = 0.5 and f = 0.25, respectively. The value f = 0.5 is presented since it is a commonly used value of the window width for Loess (it is the default value for both Minitab and S-Plus). F = 0.25 is considered since it is a more appropriate value of the parameter based on the knowledge of significant curvature in the true underlying model.

Figure 9.8 is a plot of the fit obtained using M-Regression. It closely resembles that of Figure 9.1 (as expected), the purely quadratic model with no misspecification. Table 9.1 offers the emse values calculated for M-Regression, Loess, and ORMRR as a quantitative comparison criteria of how these procedures performed on this particular data set. The emse for ORMRR is less than half that of the closest competitor (Loess with f = 0.25).

Table 9.1 EMSE values for procedures for the model in (9.1) with $\gamma = 0.75$ and error distribution CN(.10, 1.0, 5.0).

Procedure	**MREG**	**Loess** (f = 0.5)	**Loess** (f = 0.25)	**ORMRR**
emse	4.1175	7.5849	2.7598	1.3353

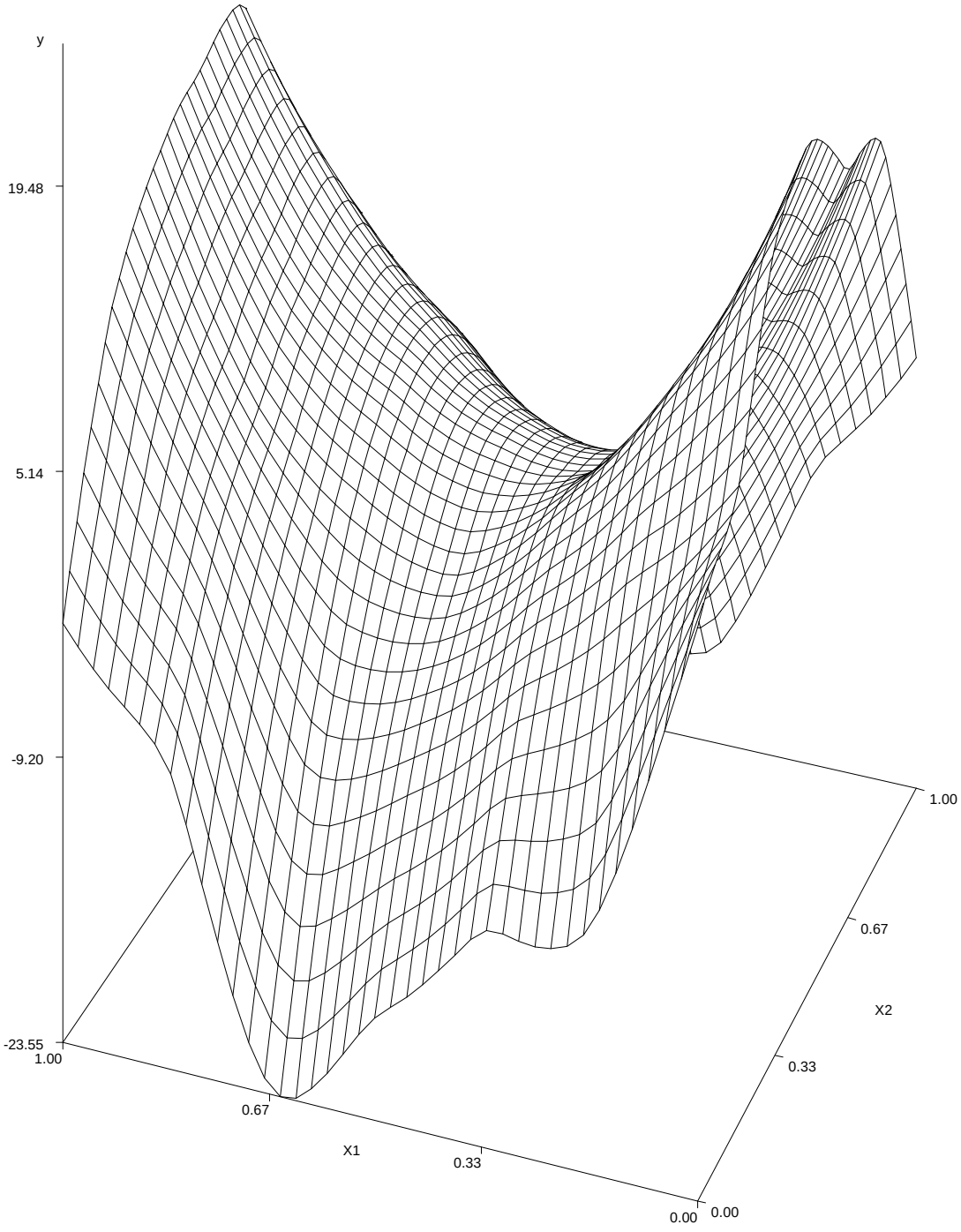

Figure 9.4 Plot of final ORMRR fit to data generated from the model $E(y|\mathbf{x}) = 2x_1^2 - 2x_2^2 + 4\gamma \cdot \sin(.75x_1 x_2)$ for $\gamma = 0.75$.

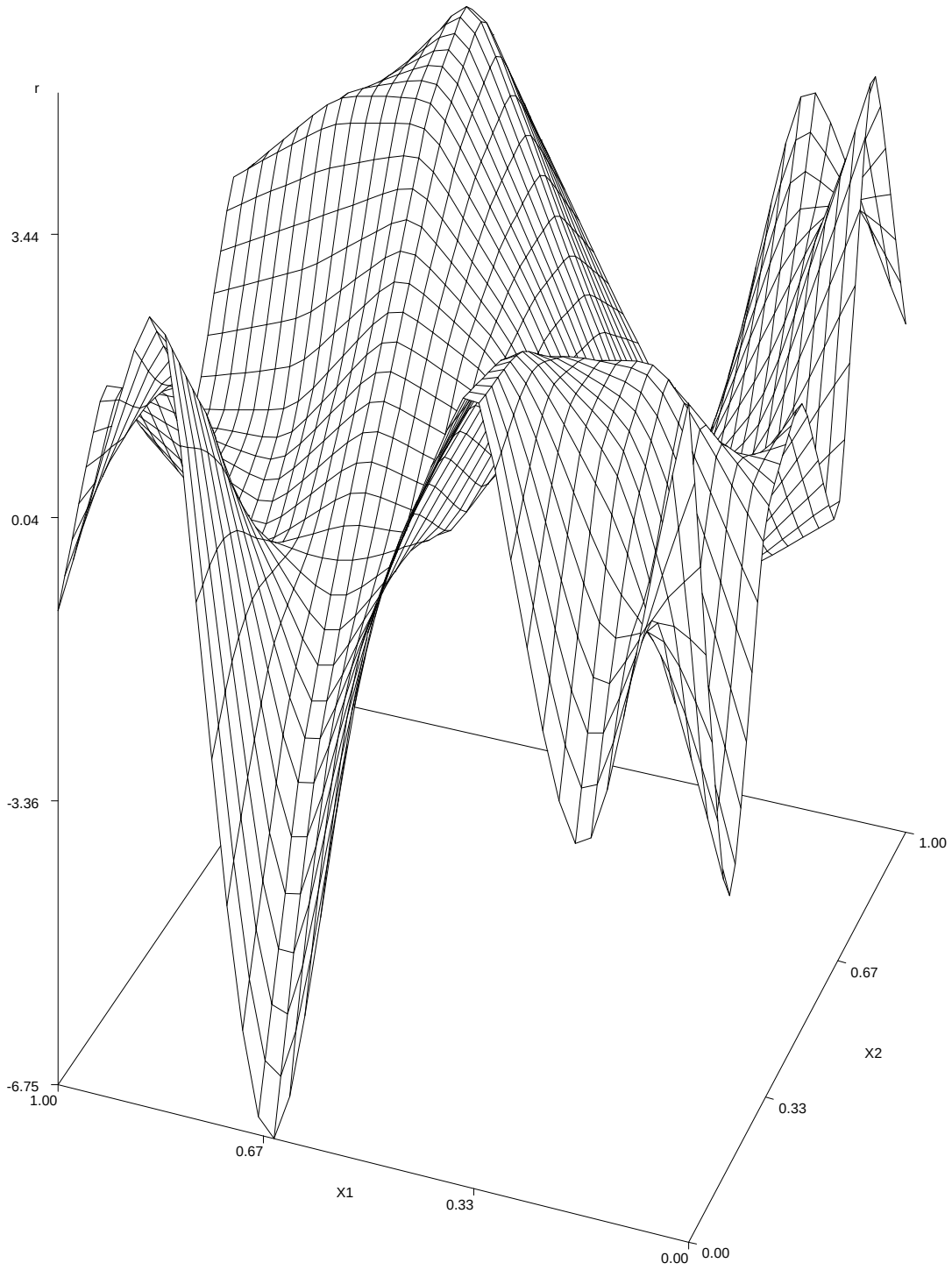

Figure 9.5 Plot of smoothed residuals that resulted from M-Regression fit to data generated from the model $E(y|\mathbf{x}) = 2x_1^2 - 2x_2^2 + 4\gamma \cdot \sin(.75x_1 x_2)$ for $\gamma = 0.75$.

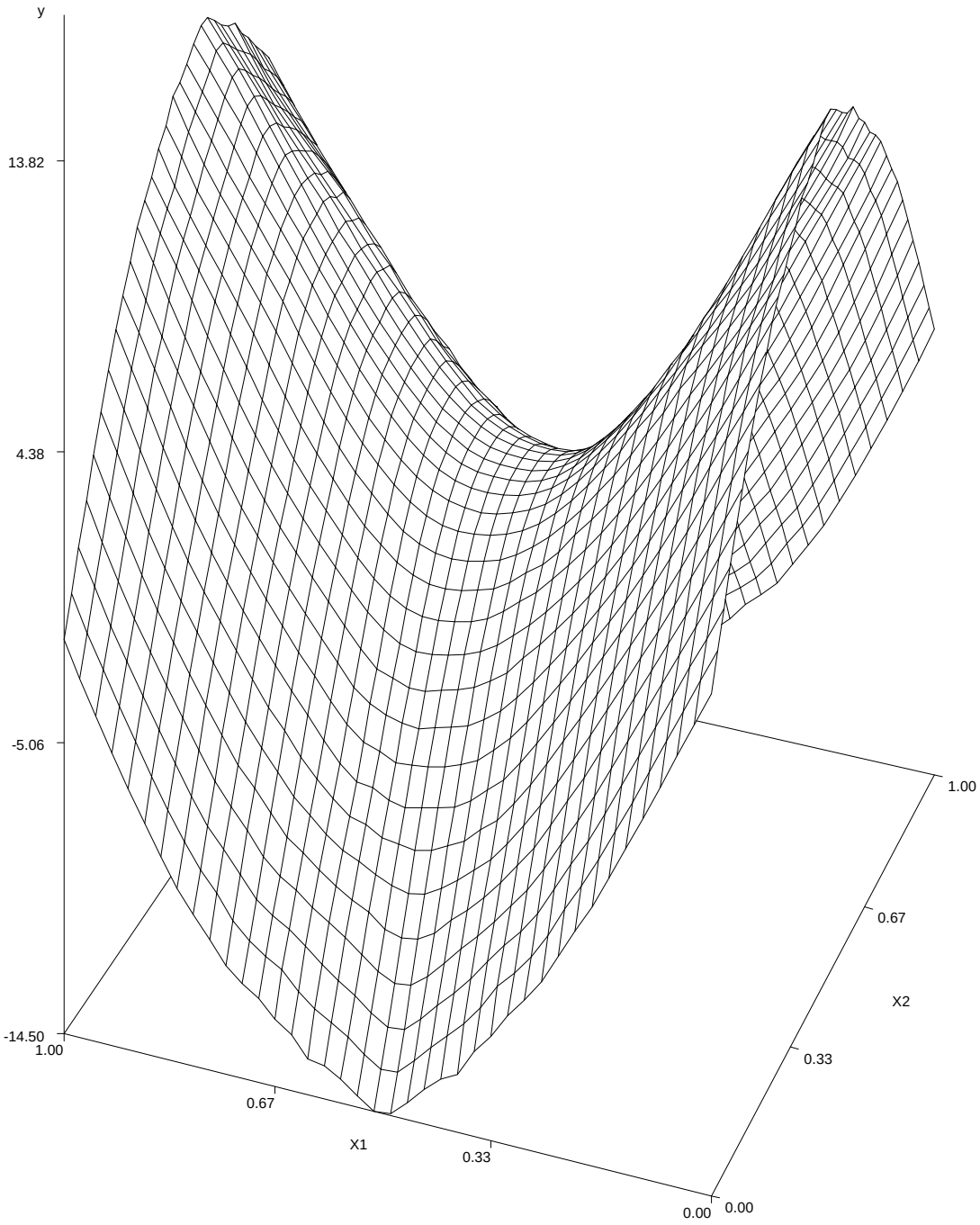

Figure 9.6 Plot of Loess fit (f = 0.5) to data generated from the model $E(y|\mathbf{x}) = 2x_1^2 - 2x_2^2 + 4\gamma \cdot \sin(.75x_1 x_2)$ for $\gamma = 0.75$.

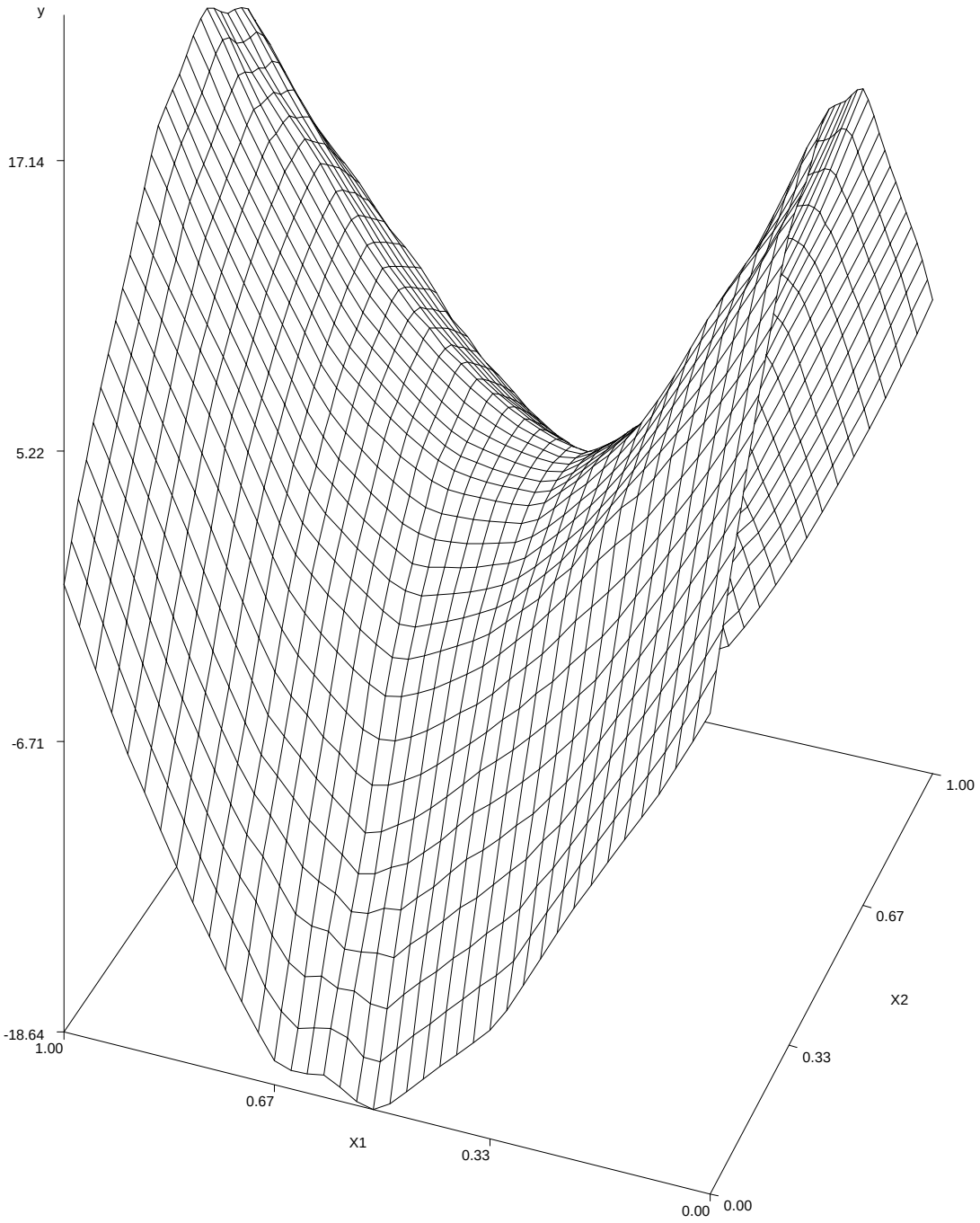

Figure 9.7 Plot of Loess fit (f = 0.25) to data generated from the model $E(y|\mathbf{x}) = 2x_1^2 - 2x_2^2 + 4\gamma \cdot \sin(.75x_1 x_2)$ for $\gamma = 0.75$.

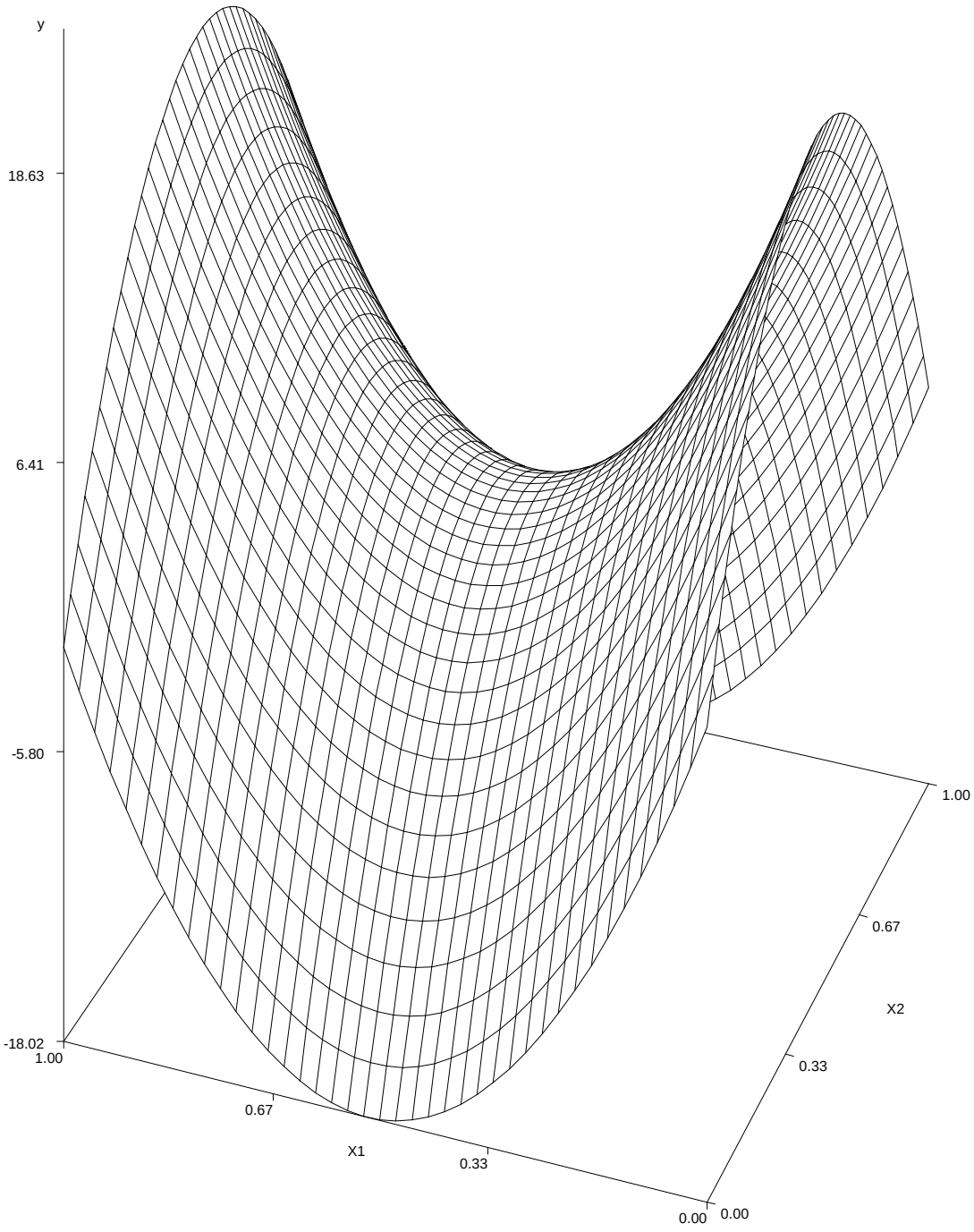

Figure 9.8 Plot of M-Regression fit to data generated from the model $E(y|\mathbf{x}) = 2x_1^2 - 2x_2^2 + 4\gamma \cdot \sin(.75x_1 x_2)$ for $\gamma = 0.75$.

As mentioned in the introduction, this example is based only on one data set and is not intended as a complete study of the behavior of these estimators in a multivariate setting. It does provide, however, a motivation for further research in this area for the ORMRR estimator, as it performs extremely well for this example. One expectation that we have is that as k, the number of regressors, increases, so does the contribution to stability of the parametric fit to the overall ORMRR fit in the multi-dimensional space. This stability, combined with the versatility of the nonparametric fit to the residuals, offers an overall fit that appears to have excellent mse properties when compared with purely parametric and purely nonparametric fits.

Something that should be pointed out is the reliance of nonparametric procedures (and thus ORMRR since it incorporates the nonparametric technique RLLR) suffer when points are sparse. This is due to the local nature of the fits, and if there are no data in a certain location, then there is no information for the nonparametric procedure. Thus, we expect ORMRR, Loess, and RLLR to suffer if the data locations are not equally spaced, as in the example above. This additional complexity that is added is another facet of the multiple regressor extension of ORMRR that we hope to investigate in future research.