NISS

Analysis of Protein Activity Data by Gaussian Stochastic Process Models

N. McMillan, J. Sacks, W. Welch, and F. Gao

Technical Report Number 29
April, 1995

National Institute of Statistical Sciences
19 T. W. Alexander Drive
PO Box 14006
Research Triangle Park, NC 27709-4006
www.niss.org

Analysis of Protein Activity Data by Gaussian Stochastic Process Models

NANCY J. MCMILLAN, JEROME SACKS, WILLIAM J. WELCH,
and FENG GAO *

National Institute of Statistical Sciences
Research Triangle Park, NC 27709

1 Introduction

1.1 A Semi-parametric Regression Model

We model protein activity as a realization of a Gaussian stochastic process which depends on explanatory variables, Buffer, Ph, NaCl, Protein Concentration, Reducing Agent, Detergent, $MgCl_2$, and Temperature as well as a number of tuning parameters. Let $\mathbf{x} = (x_1, \ldots, x_8)$ represent the vector of the eight explanatory variables, and let Y be the observed protein activity. Our basic model, as described in Sacks, Welch, Mitchell, and Wynn [4] is

$$Y(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\boldsymbol{\beta} + Z(\mathbf{x}) + \epsilon. \tag{1}$$

In general, $\mathbf{f}$ could be a vector of linear-model terms (main effects, quadratic effects, interactions, etc.), with $\boldsymbol{\beta}$ the vector of corresponding coefficients. In the model fitted in Section 3, however, $\mathbf{f}^T(\mathbf{x})\boldsymbol{\beta}$ is simply $1\beta_0$, i.e., a constant or intercept term. In our fitted model, then, systematic dependence of Y on $\mathbf{x}$ is captured by the $Z(\mathbf{x})$ term, assumed to be a Gaussian stochastic process. The ϵ term represents random error from measurement, etc.

The stochastic process $Z(\mathbf{x})$ in (1) is therefore central to our model. The correlation between $Z(\mathbf{x})$ and $Z(\mathbf{w})$, the stochastic process at two design points $\mathbf{x}$ and $\mathbf{w}$, is denoted by $R(\mathbf{w}, \mathbf{x})$.

*McMillan and Gao are Post-Doctoral Fellows and Sacks is Director, National Institute of Statistical Sciences, P.O. Box 14162, Raleigh, NC 27606. Welch is Professor, Department of Statistics and Actuarial Science, University of Waterloo, Ontario, Canada, N2L 3G1.

Specific choices for R are suggested in Section 1.2. The basic idea is that when the distance between $\mathbf{x}$ and $\mathbf{w}$ in the space of explanatory variables is small, R should be close to one; i.e., nearby values of Z, and hence the underlying function, are similar. Conversely, as $\mathbf{x}$ and $\mathbf{w}$ become more remote from each other, their responses should be unrelated, and R should approach zero. It is this property that enables Y to be predicted from the experimental data. The presence of unordered categorical variables requires that we adapt the notion of "near" versus "remote" vectors of explanatory variables. This is taken up in Section 1.2.

Let $\mathbf{x}_i$ denote the vector of explanatory variables for run i of the experiment, and let $\mathbf{Y} = (Y_1, \ldots, Y_n)^T$ denote the vector of corresponding protein activities. The model (1) can be written in matrix notation as

$$\mathbf{Y} = \mathbf{F}\boldsymbol{\beta} + \mathbf{Z} + \boldsymbol{\epsilon}, \tag{2}$$

where $\mathbf{F}$ is the expanded design matrix with $\mathbf{f}^T(\mathbf{x}_i)$ in the ith row, $\mathbf{Z} = (Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n))^T$ is the vector of stochastic process values at the n experimental settings, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T$ is the vector of random errors. Specifically, we assume that

$$\mathbf{Z} \sim \mathrm{N}(\mathbf{0}, \sigma_Z^2 \mathbf{R}), \tag{3}$$

where the $n \times n$ matrix $\mathbf{R}$ has $R(\mathbf{x}_i, \mathbf{x}_j)$ as the (i, j) element, and

$$\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma_\epsilon^2 \mathbf{I}). \tag{4}$$

Also, $\mathbf{Z}$ and $\boldsymbol{\epsilon}$ are assumed to be independent. Together, these assumptions imply that

$$\mathbf{Y} \sim N(\mathbf{F}\boldsymbol{\beta}, \sigma^2 \mathbf{C}), \tag{5}$$

where $\sigma^2 = \sigma_Z^2 + \sigma_\epsilon^2$, and the $n \times n$ correlation matrix $\mathbf{C}$ is given by $(\sigma_Z^2 \mathbf{R} + \sigma_\epsilon^2 \mathbf{I})/\sigma^2$.

As expressed in (5) this model can easily be recognized as a *universal kriging* model, common to the spatial statistics literature [3], as well as to the analysis of deterministic experiments such as those discussed in [1], [2], [4], [6], and [7] on which our work is based. This model has proven to be useful in situations of both types, specifying flexible nonlinear models. In the kriging scenario, explanatory variables are actual physical locations in two or three dimensions, and the covariance structure, R, is specified via a variogram rather than a covariance matrix.

When R and σ_Z^2/σ^2 are assumed known, we use the best linear unbiased predictor (BLUP) of $Y(\mathbf{x})$,

$$\hat{Y}(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\hat{\boldsymbol{\beta}} + \mathbf{c}(\mathbf{x})^T \mathbf{C}^{-1}(\mathbf{Y} - \mathbf{F}\hat{\boldsymbol{\beta}}). \tag{6}$$

Here $\mathbf{c}(\mathbf{x})$ is a vector with element i given by $\frac{\sigma_Z^2}{\sigma^2}R(\mathbf{x}, \mathbf{x}_i)$, the correlations between the Y's at $\mathbf{x}$ and the n experimental runs. The Gaussian assumption of (5) implies that (6) is also the maximum likelihood estimate of $Y(\mathbf{x})$. From a Bayesian viewpoint, under model (5) and a non-informative prior on $\boldsymbol{\beta}$, (6) is the posterior mean for prediction.

1.2 Specifying the Covariance Structure

Model (5) has previously been employed primarily for *continuous* or at least *ordinal* explanatory variables. We propose a modification to the structure of R to handle unordered *categorical* variables.

In previous work [4] with q quantitative (continuous or ordinal) explanatory variables, the correlation $R(\mathbf{w}, \mathbf{x})$ between $Z(\mathbf{x})$ and $Z(\mathbf{w})$ was specified as

$$R(\mathbf{w}, \mathbf{x}) = \prod_{i=1}^{q} \exp\{-\theta_i |w_i - x_i|^{p_i}\}, \tag{7}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)$ and $\mathbf{p} = (p_1, \ldots, p_q)$ are unknown parameters. All p_i's are in $(0, 2]$ in order to satisfy a positive definiteness constraint. The θ_i's are restricted to be nonnegative, implying that the correlation is nonincreasing with distance in each dimension. Unlike most traditional kriging models, no isotropic assumptions have been made.

Suppose x_i is an unordered categorical variable with k levels. In this case we replace the factor $\exp\{-\theta_i |w_i - x_i|^{p_i}\}$ in (7) by

$$\prod_{j=1}^{k} \exp\{-\theta_{ij} |I(w_i = j) - I(x_i = j)|\} \tag{8}$$

where $I(x_i = j)$ is 1 if x_i takes level j and 0 otherwise.

For example, buffer is categorical with four levels. Ignore the other explanatory variables, so that we may omit the i subscript in θ_{ij}, and let $Z(j)$ denote the stochastic process at buffer j. The correlations between $Z(1), \ldots, Z(4)$ are

$$\mathrm{Corr}(Z \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}) = \begin{pmatrix} 1 & & & \\ \exp\{-(\theta_1 + \theta_2)\} & 1 & & \\ \exp\{-(\theta_1 + \theta_3)\} & \exp\{-(\theta_2 + \theta_3)\} & 1 & \\ \exp\{-(\theta_1 + \theta_4)\} & \exp\{-(\theta_2 + \theta_4)\} & \exp\{-(\theta_3 + \theta_4)\} & 1 \end{pmatrix}. \tag{9}$$

For four or more levels, this correlation matrix is restricted (for $k = 4$, six correlations are specified by four θ_{ij} parameters). For three levels, the matrix is fully parameterized, while for $k = 2$ it is over-parameterized. The latter problem is easily avoided by setting one of the two θ_{ij} parameters to zero.

2 Model Fitting and Analysis

Model (5) is parameterized by $\boldsymbol{\beta}$, σ_Z^2, σ_ϵ^2, $\boldsymbol{\theta}$, and $\mathbf{p}$, which we estimate by maximum likelihood estimation (MLE) [4]. The estimates of $\boldsymbol{\beta}$ and σ^2, given $\boldsymbol{\theta}$, $\mathbf{p}$, and σ_Z^2/σ^2, are

$$\hat{\boldsymbol{\beta}} = (\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{C}^{-1}\mathbf{Y} \tag{10}$$

and

$$\hat{\sigma}^2 = \frac{1}{n}(\mathbf{Y} - \mathbf{F}\hat{\boldsymbol{\beta}})^T\mathbf{C}^{-1}(\mathbf{Y} - \mathbf{F}\hat{\boldsymbol{\beta}}). \tag{11}$$

Estimates of $\boldsymbol{\theta}$, $\mathbf{p}$, and σ_Z^2/σ^2 must be obtained by numerical minimization of $|\mathbf{C}|^{1/n}\hat{\sigma}^2$.

While performance of the kriging predictor (6) is of primary interest, identifying the important explanatory variables aids interpretation. For designed experiments on a cuboidal region, it is useful to define the mean effect by $\mu_0 = \int Y(\mathbf{x}) \prod dx_h$, the main effect of x_i by $\mu_i(x_i) = \int Y(\mathbf{x}) \prod_{h \neq i} dx_h$, and the joint effect of x_i and x_j by $\mu_{ij}(x_i, x_j) = \int Y(\mathbf{x}) \prod_{h \neq i,j} dx_h$. We estimate these effects by replacing $Y(\mathbf{x})$ with (6). Plotting the estimated effects gives a visual indication of the relative importances of the factors.

The performance of the kriging predictor (6) can be judged by leave-one-out cross-validation. The cross-validation predictor of $Y(x_i)$ based on all data except run i is

$$\hat{Y}_{-i}(\mathbf{x}_i) = \mathbf{f}^T(\mathbf{x}_i)\hat{\boldsymbol{\beta}}_{-i} + \mathbf{c}_{-i}^T\mathbf{C}_{-i}^{-1}(\mathbf{Y}_{-i} - \mathbf{F}_{-i}\hat{\boldsymbol{\beta}}_{-i}) \tag{12}$$

where

$$\hat{\boldsymbol{\beta}}_{-i} = (\mathbf{F}_{-i}^T\mathbf{C}_{-i}^{-1}\mathbf{F}_{-i})^{-1}\mathbf{F}_{-i}^T\mathbf{C}_{-i}^{-1}\mathbf{Y}_{-i}. \tag{13}$$

We use the subscript $-i$ to denote the deletion of the ith row of $\mathbf{c}$, $\mathbf{F}$, and $\mathbf{Y}$, and both the ith row and the ith column of $\mathbf{C}$.

We perform two versions of this cross validation. In the first, the parameters $\boldsymbol{\theta}$, $\mathbf{p}$, and σ_Z^2/σ^2 determining the correlation structure (hence $\mathbf{c}$ and $\mathbf{C}$) are estimated from all the data. In the second version, these correlation parameters are re-estimated by MLE at each deletion (i.e., without the deleted run). Previous work (for example [6]) suggested that the first method gives a valid indication of prediction accuracy, obviating the second, computationally more demanding, version. For the model fitted in Section 3, however, the second version is apparently necessary.

To investigate the fit of the model we can also compute the "theoretical" MSE of $\hat{Y}_{-i}(\mathbf{x}_i)$, assuming the correlation structure fixed and known,

$$\mathrm{MSE}[\hat{Y}_{-i}(\mathbf{x}_i)] = \sigma^2 \left[1 - (\mathbf{f}(\mathbf{x}_i)^T, \mathbf{c}_{-i}(\mathbf{x}_i)^T) \begin{pmatrix} \mathbf{0} & \mathbf{F}_{-i}^T \\ \mathbf{F}_{-i} & \mathbf{C}_{-i} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{f}(\mathbf{x}_i) \\ \mathbf{c}_{-i}(\mathbf{x}_i) \end{pmatrix} \right]. \tag{14}$$

As with the cross-validation predictions themselves, this MSE can be based on correlation parameters from the full data or re-estimated after deleting run i. When prediction variability, as indicated by (14), is large relative to the range of the data, suspicion is raised about the utility of the model being fit. Using the square root of MSE$[\hat{Y}_{-i}(\mathbf{x}_i)]$ for the standard errors to normalize the cross-validation prediction residuals should result in standard normal errors. This can be checked by a Q-Q plot and provides a rough check of the validity of the Gaussian modelling assumptions.

We also evaluate the model based on an empirical measure of MSE averaged over the design points,

$$\text{MSE} = \frac{1}{n - \text{tr}(\mathbf{H})} \sum_{i=1}^{n} (\hat{Y}(\mathbf{x}_i) - Y(\mathbf{x}_i))^2. \tag{15}$$

The "hat" matrix, $\mathbf{H}$, is defined by

$$\mathbf{H} = (\mathbf{I} - (\sigma_z^2/\sigma^2)\mathbf{R}\mathbf{C}^{-1})\mathbf{F}(\mathbf{F}^T\mathbf{C}^{-1}\mathbf{F})^{-1}\mathbf{F}^T\mathbf{C}^{-1} + (\sigma_z^2/\sigma^2)\mathbf{R}\mathbf{C}^{-1}. \tag{16}$$

We use the trace of the "hat" matrix as a surrogate for the degrees of freedom in the model as suggested by Wahba [5] for splines.

3 Protein Activity Model

The specific model we employed for the protein activity data included only a constant term, β_0, in the regression part. The correlation structure was specified with 17 θ's and four p's: There is one (θ, p) pair for each of the four continuous variables, and one θ for each level of the four categorical variables. In addition to these 21 parameters, β_0, σ_Z^2, and σ_ϵ^2 are also estimated by MLE. Many of the θ's are estimated to be zero. This is "encouraged" by the algorithm by periodically testing whether setting each θ parameter equal to zero produces an insignificant change in the likelihood. If θ_i is zero for a continuous variable or the θ_{ij}'s are all zero for a categorical variable, that variable has no effect on the predictor. This provides a "screening" for important effects [6].

Of the 21 correlation parameters, 12 parameters are "active." These are the θ_{ij}'s for Buffer TRS, Buffer P04, Reducing Agent AGX, Detergent TWEEN, Detergent N_OCTGLU, and $\text{MgCl}_2 = 1$, plus the θ_i's and p_i's for NaCl, Protein Concentration, and Temperature. All of the active p_i's were estimated to be 2 indicating smoothness (differentiability) of the response surface. The constant term, β_0, was estimated to be 0.7664, model variance, σ_Z^2, was estimated at 0.438, and error variance, σ_ϵ^2, was estimated at 0.00708. The maximum likelihood estimate of the standard deviation

Table 1: Non-zero parameter estimates. (All p's were estimated at 2.)

Parameter	Estimate	Parameter Description
β_0	0.7664	Mean
σ_Z^2	0.438	Model variance
σ_ϵ^2	0.00708	Error variance
θ_1	0.149	TRS indicator
θ_2	1.07	P04 indicator
θ_3	0.000000337	NaCl
θ_4	0.000000658	Protein Concentration
θ_5	0.0692	AGX indicator
θ_6	11.8	TWEEN indicator
θ_7	0.579	N_OCTGLU indicator
θ_8	0.0814	$MgCl_2$
θ_9	0.000407	Temperature

of error, $\sigma_\epsilon = 0.0841$, agreed rather well with the estimate of measurement error based on the repeated design point in the data, 0.0965.

Not surprisingly, in light of the estimate of σ_ϵ^2, the plot in Figure 1 of the predicted values based on (6) versus the observed data showed an excellent fit. The trace of $\mathbf{H}$ is approximately 81 which explains the approximate interpolation of the model. RMSE, calculated by taking the square root of (15), is 0.0877. Our leave-one-out cross-validation (12) based on the economical method of *not* re-fitting correlation parameters at each deletion resulted in a CVrootMSE of 0.3187. The plot in Figure 2 of the cross-validation predictions from this method against the observed data shows larger scatter than Figure 1, but no systematic problems. The "theoretical" prediction errors (14) are large relative to the range of the data. The "theoretical" root mean squared errors, given by the square root of (14), are roughly between 0.1 and 0.5. Since the random error component in this model is small, this suggests that the variability due to the model is large, encouraging further investigation about the model including the re-estimation of the correlation parameters at each deletion in computing CVrootMSE. The Q-Q plot in Figure 3, based on economical cross-validation (12), of the standardized residuals from cross-validation seemed to validate the Gaussian assumption.

Our analysis of the main and joint effects produced three important main effects, Protein Concentration, Buffer, and Detergent, and four important joint (interaction) effects, Buffer × Detergent, Buffer × Temperature, Detergent × Protein Concentration, and Detergent × Temperature. As each of the important main effects appears at least once in an important joint effect, we need only look at the joint effects in Figure 4 to find optimal conditions. In Figure 4 the lines for Buffer MES and Buffer HPS are indistinguishable; the same is true for Detergents ETHGLY and GLYCINE.

Our conclusions from Figure 4 are somewhat mixed. It is clear that optimal conditions include Detergent TWEEN and that high Temperature is at least somewhat better than low Temperature. Choice of an optimal Buffer is clouded by the conflicting evidence of the Buffer × Detergent table and the Buffer × Temperature plot. The former recommends either Buffer MES or HPS while the latter encourages Buffer PO4 at a high temperature. Further investigation of the three-way interaction between Buffer, Detergent, and Temperature might resolve this conflict. Once Detergent TWEEN is chosen, Protein Concentration is fairly unimportant. The settings of Ph, NaCl, Reducing Agent, and $MgCl_2$ are also unimportant. These findings are consistent with our selection of an optimal design point (design point which maximizes our predictor, (6)) as case 37.

We performed the requested leave-one-out cross-validation, leaving out only one of the largest third of the data points each time by the two methods previously discussed. Our CVrootMSE based on only recalculating (6) without re-estimating the correlation parameters (CVPredicted_2 in Table 3) was 0.375, re-estimating the correlation parameters at each deletion (CVPrediction_1 in Table 3) gave a CVrootMSE = 0.536.

The discrepant performance of our first estimator of cross-validation error lead us to a more careful examination of the data set. We found that the surface is extremely spiky. For example, cases 77 and 82 both have Buffer HPS, Detergent N_OCTGLU, Temperature −80°, and Protein Concentration 100, i.e., they have the same settings for the apparently important variables. Yet case 77 has a high response, 1.495, while case 82 has an extremely low response, 0.032. Similar examples exist in this data set. We would be suspicious of any method which put unexplained spikes into a predictor when the only data point which seems to support that spike has been removed from the data. There were no other design points in the top 33 observations which had the same important explanatory variables (Buffer, Detergent, Temperature, and Protein Concentration) as case 77.

A cursory check of main and joint effects with case 77 deleted lead to similar recommendations on optimal values in the important effects. We recommend that further design points be chosen in the optimal areas suggested by this analysis in order to more accurately predict optimal protein storage conditions.

References

[1] M. C. Bernardo, R. Buck, L. Liu, W. A. Nazaret, J. Sacks, and W. J. Welch, *Integrated circuit design optimization using a sequential strategy*, IEEE Transactions on Computer-aided Design **11** (1992), 361–372.

[2] K. P. Bowman, J. Sacks, and Y. Chang, *Design and analysis of numerical experiments*, Journal of the Atmospheric Sciences **50** (1993), 1267–1278.

[3] Noel Cressie, *Statistics for spatial data*, John Wiley & Sons, Inc., 1991.

[4] J. Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, *Design and analysis of computer experiments*, Statistical Science **4** (1989), 409–435.

[5] G. Wahba, *Spline models for observational data*, Society for Industrial and Applied Mathematics, Philadelphi, Pennsylvania, 1990.

[6] W. J. Welch, R. J. Buck, J. Sacks, H. P. Wynn, T. J. Mitchell, and M. D. Morris, *Screening, predicting, and computer experiments*, Technometrics **34** (1992), 15–25.

[7] W. J. Welch and J. Sacks, *A system for quality improvement via computer experiments*, Communications in Statistics-Theory Methodology **20** (1991), 477–495.

Table 2: Cross-validation of 33 largest observations.

Case	Observed	Predicted	CVPredicted$_1$	CVPredicted$_2$
37	2.112	2.090123	1.202510	1.5591007
76	2.059	1.998767	0.971013	1.5342398
90	1.940	1.929449	1.766920	1.7875905
20	1.880	1.845872	0.815522	0.8106532
38	1.813	1.776779	0.770639	0.9548504
65	1.793	1.776152	1.596580	1.6685995
27	1.791	1.788682	1.746870	1.7390436
79	1.732	1.725929	1.660820	1.6715659
29	1.716	1.646411	1.179970	1.2297464
68	1.690	1.671872	1.196730	1.2518140
21	1.670	1.646340	1.240200	1.3185723
88	1.664	1.666200	1.634900	1.6846383
39	1.639	1.640562	1.865850	1.6643416
56	1.586	1.574286	2.078500	1.5134114
64	1.554	1.554969	1.580970	1.5679600
54	1.544	1.555517	1.700540	1.6636751
77	1.495	1.452651	0.183360	0.6809593
95	1.441	1.338367	1.287620	1.2884319
3	1.400	1.398416	1.336880	1.3475908
69	1.378	1.355281	1.381460	1.2516411
67	1.343	1.312942	0.138577	0.7787523
89	1.328	1.343454	0.852592	1.6695348
96	1.309	1.338367	1.353340	1.3526548
1	1.253	1.338367	1.380280	1.3799009
94	1.230	1.208312	0.803463	0.9498886
92	1.201	1.201713	1.213300	1.2178295
40	1.158	1.166602	1.269750	1.2479975
55	1.134	1.074884	1.019610	1.0204219
47	1.084	1.153612	1.661780	1.5061271
86	1.049	1.052411	1.092090	1.0912879
17	1.039	1.024496	0.809032	0.8270347
35	1.016	1.074884	1.115570	1.1291317
13	1.000	1.008796	1.313500	1.2592392

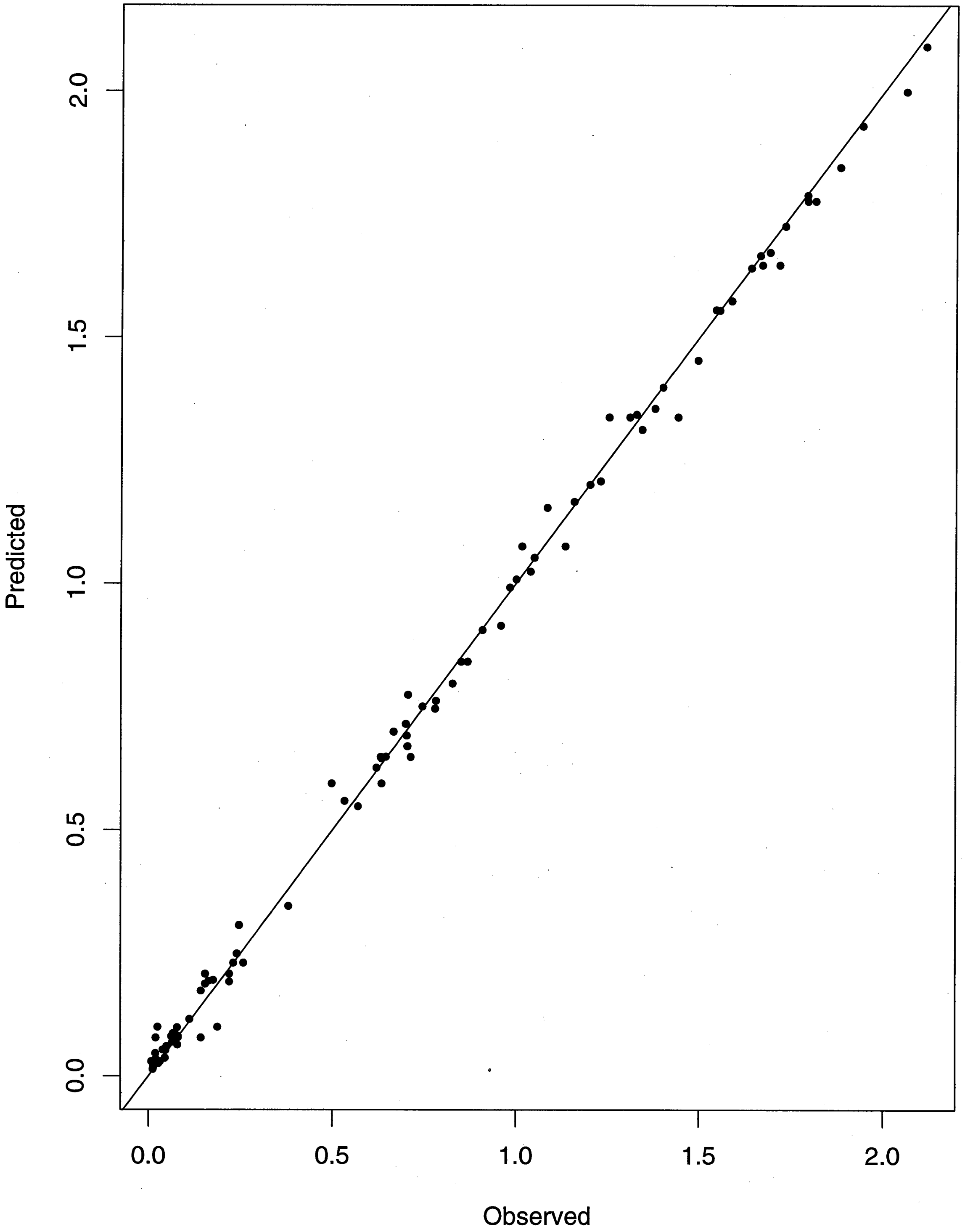


Figure 1: Plot of values predicted from model against observed responses.

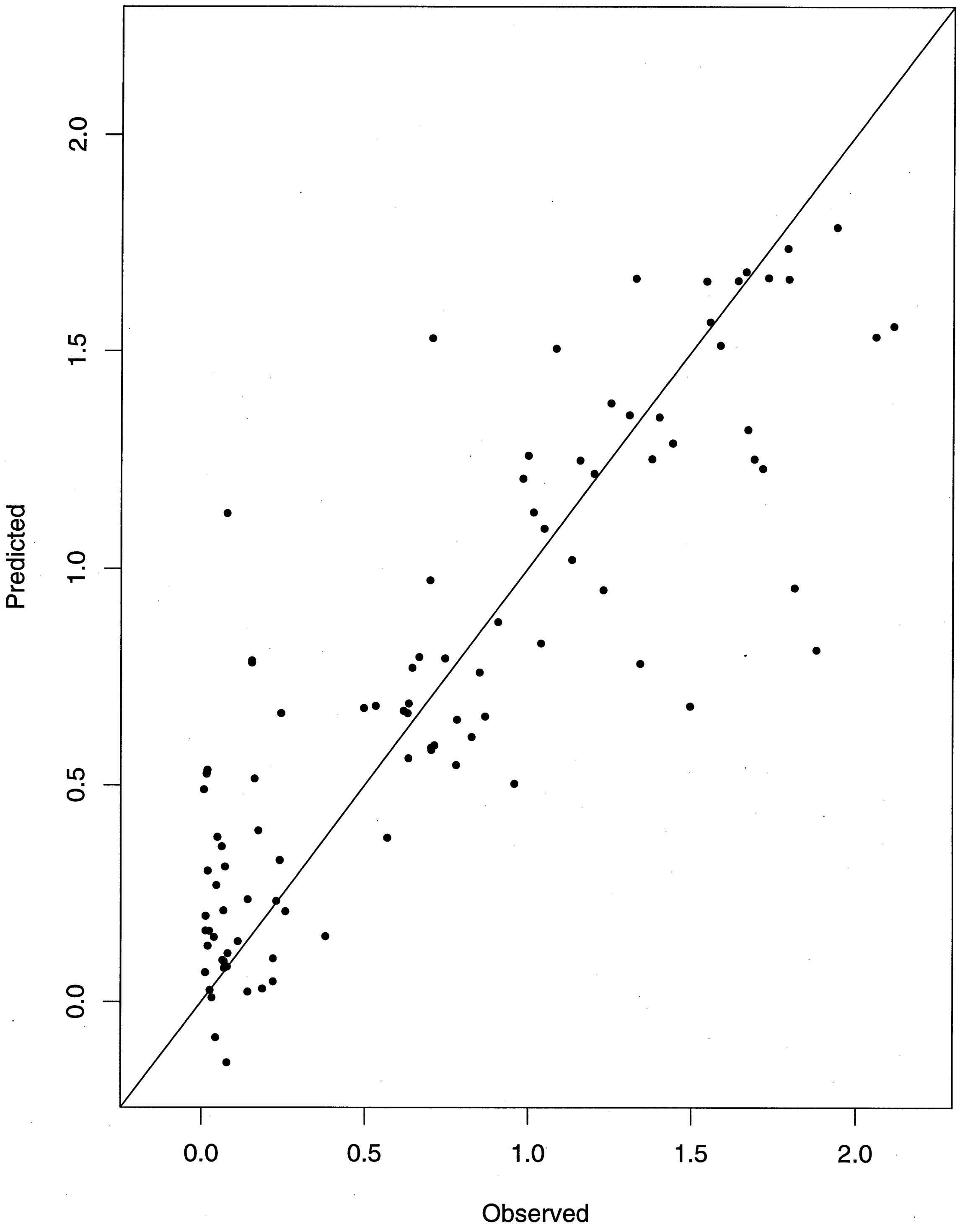


Figure 2: Plot of values predicted by cross-validation against observed responses.

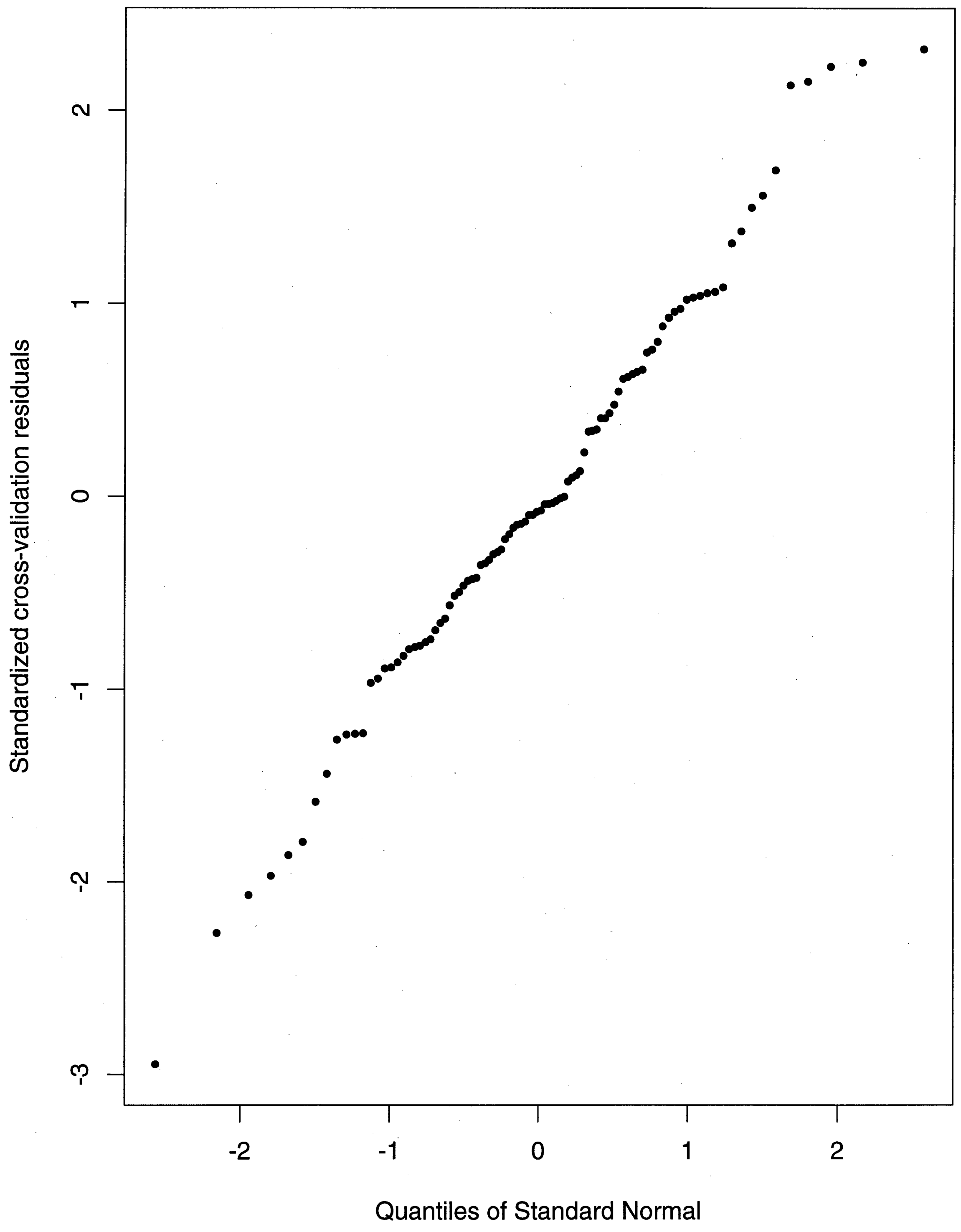
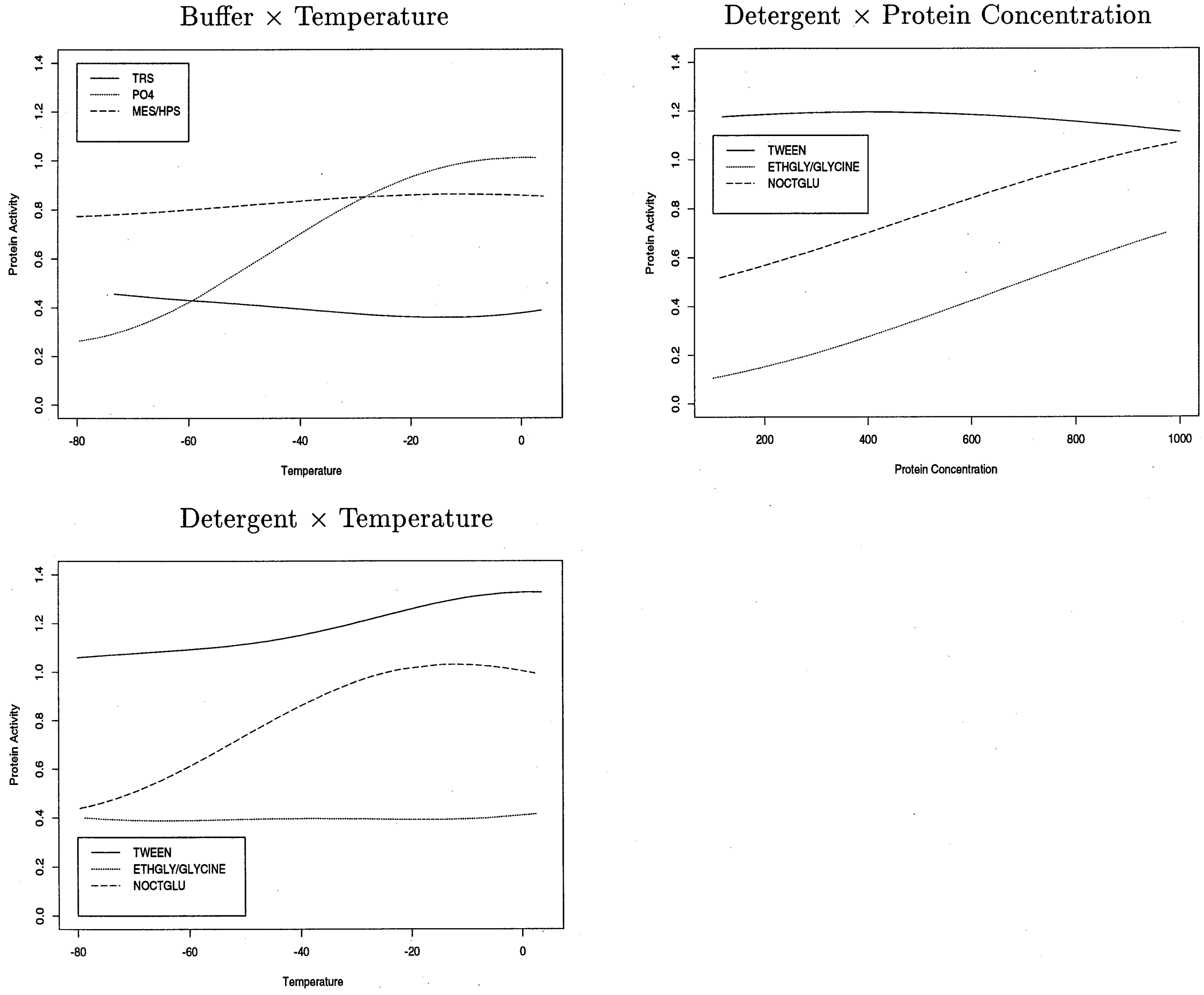


Figure 3: Q-Q plot of standardized residuals from cross-validation.

Buffer × Detergent

	TWEEN	*ETHGLY*	*GLYCINE*	*N_OCTGLU*
TRS	1.03	0.27	0.27	0.05
*P*04	0.97	0.45	0.45	0.93
MES	1.36	0.43	0.43	1.14
HPS	1.36	0.43	0.43	1.14



Buffer × Temperature

Detergent × Protein Concentration

Detergent × Temperature

Figure 4: Important joint effects.