Optimization of Semiconductor Monitoring via Residual-Based Computational and Statistical Methods

Ulduz Mammadova

Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Alanya University, Antalya, Turkey
contact@ulduzmammadova.com

Abstract. Automated detection of process deviations in semiconductor manufacturing is crucial for ensuring product quality and operational efficiency. Industrial datasets, such as the dataset, are challenging because they are relatively high-dimensional, contain highly correlated features, and include binary outcomes. This study develops a computational framework for residual-based monitoring, combining statistical modeling with algorithmic techniques to detect deviations.

For the monitoring procedure, the data set was split evenly: the first half was used to create baseline models and define control limits, while the remaining half served for testing. Principal Component Analysis and Ridge regression are employed to minimize the negative impacts of multicollinearity. Principal Component Analysis generates linear transformations of the features, reducing the dimensionality of the data while preserving most of its variation. Ridge regression, on the other hand, preserves all features and applies a penalty to large coefficients while maintaining the relationship between predictors and the response. In both scenarios, deviance residuals and randomized quantile residuals (RQRs) were computed. These residuals are proven to be approximately normally distributed, which enables accurate monitoring.

The actual monitoring was conducted using traditional control charts, median absolute deviation (MAD) limits, and kernel density estimation (KDE)–based limits. Among these techniques, traditional charts are straightforward and interpretable, MAD limits reduce the influence of extreme residuals, and KDE leverages non-parametric density estimation to capture subtle process deviations. These methods integrate algorithmic computation, statistical reasoning, and automated evaluation to establish a foundation for detecting process abnormalities.

The results indicate that KDE-based monitoring detects deviations more quickly and flags a larger number of observations than the other methods. Traditional and MAD-based approaches produce similar but slightly less sensitive results.

Keywords: Anomaly Detection, Kernel Density Estimation, Ridge regression, Principal Component Analysis, Semiconductor Manufacturing.

R. Rzayev et al. (eds.), *Proceedings of the International Conference on Current Problems in Engineering and Applied Sciences (ICCPEAS 2025)*, Advances in Engineering Research 299,
https://doi.org/10.2991/978-94-6239-668-5_87

1 Introduction

Consistent surveillance of semiconductor manufacturing through the collection of signals from sensors and process measurement points allows close monitoring of the process. By treating each sensor's information as a feature and the output as a response variable, such production data can be analyzed for anomalies using modeling. Unfortunately, not all datasets are suitable for straightforward modeling. Industrial datasets, in particular, tend to be more challenging due to their complex nature. They can be relatively high-dimensional, contain highly correlated features, and the outcome does not always follow a normal distribution. When the response belongs to the exponential family (binomial, Poisson, gamma, etc.), generalized linear models (GLMs) are used to represent the relationship between features and the response variable [1].

Traditional methods of model parameter estimation mainly rely on independence among feature variables. In the case of interrelated features, estimating model parameters becomes particularly challenging. One approach to overcome this limitation is the application of Principal Component Analysis (PCA). PCA performs a transformation that reduces the dimensionality of the dataset by creating linear combinations of the original features, capturing most of the variance while simplifying the feature space [2]. An alternative to PCA is a penalized estimation approach. In contrast to PCA, penalized estimation methods like ridge regression keep all original features but impose a penalty for large coefficients. It stabilizes estimates and preserves the predictive relationship between the predictors and the response. Initially, [3] proposed a ridge estimator for GLMs. Researchers later extensively studied the ridge estimator for logistic regression [4-7].

A well-fitted model provides a reliable baseline for monitoring, as residuals computed from the model capture the differences between observed and predicted responses, effectively summarizing the entire dataset in a single variable. Several authors [8-14] have used residuals in combination with traditional control charts for process monitoring. Normality of the residuals enables easy monitoring of such processes. [15] showed that deviance residuals calculated from a fitted GLM follow an asymptotic normal distribution, while [16] proved the normality of randomized quantile residuals under a true GLM through a series of simulation studies.

In this study, we evaluated three types of monitoring charts based on both deviance and RQR residuals for semiconductor monitoring data. The dataset is retrieved from a publicly available source and subjected to monitoring. Monitoring is carried out using the Shewhart chart, considering the process mean, the MAD-based chart, and the KDE-based chart. The MAD chart is adopted to overcome the sensitivity of classical charts to extreme residuals, providing a more robust measure of process variation [17]. The KDE chart represents an entirely different approach, relying on non-parametric density estimation to capture deviations in the overall distribution, including shifts in location, changes in spread, and subtle shape anomalies [18].

The paper is structured as follows: Section 2 discusses modeling options for data with binary outcomes. Section 3 describes the monitoring methods. Section 4 describes the dataset, as well as the details about the monitoring technique and findings. Section 5 concludes the paper.

2 Modeling Strategies

Let $X_{n\times p}$ be a feature matrix and $y_{n\times 1}$ binary vector of the response variable. The relationship between features and response can be modeled using a logistic regression model [1]. Logistic regression is a special case of the GLM where the relationship between features and response is modeled using a link function; in our case, it is a logit link function as

$$\pi_i = P(y_i = 1|x_i) = \frac{\exp(x_i\beta)}{1 + \exp(x_i\beta)}, \qquad for\ i = 1,2,\dots,n$$

or equivalently

$$\log\left(\frac{\pi}{1-\pi}\right) = x_i\beta, \qquad for\ i = 1,2,\dots,n.$$

Here, x_i donates the i-th row vector in the feature matrix X, and $\beta = (\beta_1, \beta_2, \dots, \beta_p)'$ is the vector of unknown model parameters.

The parameter vector β can be estimated using several approaches, one of which is maximum likelihood estimation (MLE). Since a closed-form solution is not available, the Iteratively Reweighted Least Squares (IRLS) algorithm – based on the Newton–Raphson method is used to compute the MLEs. Following [19], the IRLS update for logistic regression is given by

$$\hat{\beta} = \left(X'\widehat{W}X\right)^{-1}X'^{\widehat{W}}z \tag{1}$$

where $z = X\hat{\beta} + \widehat{W}^{-1}(y - \hat{\pi})$ and $\widehat{W}_{n\times n}$ is a diagonal weight matrix with elements $\widehat{w}_{ii} = n_i\hat{\pi}_i(1-\hat{\pi}_i), i = 1,2,\dots,n$.

Even though classical estimation is reliable, these methods may become unstable when the predictors are highly correlated. To address multicollinearity, PCA can be applied to transform the original features into a set of uncorrelated components. This transformation can be efficiently performed using Singular Value Decomposition (SVD), which factorizes the matrix X as

$$X = U\Sigma V'$$

where $U_{n\times p}$ is the matrix of left singular vectors, $\Sigma_{p\times p}$ is the diagonal matrix of singular values, and $V_{n\times p}$ is the matrix of right singular vectors. The first s column of the $U\Sigma$ is the principal components retained in the model, capturing the majority of variance in the dataset. By projecting the data onto these components, a transformed uncorrelated feature matrix X_s^* can be obtained. Substituting X by X_r^* in Eq. (1) gives

$$\hat{\beta}_r = \left(X_s^{*\prime}\widehat{W}X_s^*\right)^{-1}X_s^{*\prime}\widehat{W}z^*$$

where $z^* = X_s^* \hat{\beta} + \widehat{W}^{-1}(y - \hat{\pi})$. The final coefficient estimates in the original feature space are recovered as $\hat{\beta} = V_s \hat{\beta}_s$ where V_s contains the first s column of V.

An alternative approach for reducing the influence of correlated predictors is to apply penalized estimation, such as ridge regression, which is also suitable for logistic models. The ridge-penalized objective function is given by

$$\mathcal{L} = -\sum_{i=1}^{n} [y_i(x_i\beta) - \log(1 - \exp(x_i\beta))] + k\sum_{j=1}^{p} \beta_j^2$$

where k is the shrinkage parameter. The corresponding Ridge estimator is

$$\hat{\beta}_k = \left(X'\widehat{W}_k X + kI_p\right)^{-1} X'\widehat{W}_k z_k$$

with $z_k = X\hat{\beta}(k) + \widehat{W}_k^{-1}(y - \hat{\pi})$ and $\widehat{W}_k$, the diagonal weight matrix constructed analogously to $\widehat{W}$.

In practice, some packages utilize optimization algorithms for parameter estimation. The "`glmnet`" [20] package available in R uses a cyclical coordinate descent algorithm to estimate the ridge-penalized logistic regression coefficients. This algorithm cycles through all predictors until convergence, updating each coefficient iteratively while keeping the others constant.

2.1 Residuals

According to [7], the GLM framework, deviance residuals for the logistic regression model can be determined as

$$d_i = sign(y_i - \hat{\pi}_i)\sqrt{2\left[y_i \log\left(\frac{y_i}{\hat{\pi}_i}\right) + (1 - y_i)\log\left(\frac{y_i}{1 - \hat{\pi}_i}\right)\right]}, i = 1,2,\dots,n$$

Following [16] for binary response, the randomized quantile residual for the i-th observation can be defined as

$$rq_i = \Phi^{-1}\left(F(y_i^-; \hat{\pi}_i) + u_i p(y_i; \hat{\pi}_i)\right), i = 1,2,\dots,n$$

where $\Phi^{-1}(*)$ is the standard normal quantile function, $F(y_i^-; \hat{\pi}_i)$ is the left-limit of the cumulative distribution function, $p(y_i; \hat{\pi}_i)$ is the probability mass function evaluated at y_i, and $u_i \sim U(0,1)$.

The functional form of the both deviance and RQR residuals remains the same regardless of the estimation approach; however, the values of $\hat{\pi}_i$ differ depending on whether the coefficients are obtained through classical maximum likelihood estimation, PCA-based logistic regression, or ridge-penalized estimation.

3 Residual-Based Monitoring

Residual-based monitoring relies on the principle that, under a well-fitted model, residuals should be approximately normally distributed. Significant departures from this expectation indicate potential process deviations or abnormalities. Residuals from logistic regression – whether deviance or RQR can be used as the basis for process monitoring.

The Shewhart chart is one of the commonly used control charts. In this method, residuals are plotted sequentially, and control limits are set based on the data from the process that is considered in an in-control state. Control limits for monitoring the mean are

$$UCL_{mean} = \hat{\mu} + l\hat{\sigma}$$
$$LCL_{mean} = \hat{\mu} - l\hat{\sigma}$$

where $\hat{\mu}$ and $\hat{\sigma}$ are the mean and standard deviation of residuals when the process is in-control, l, on the other hand, is a constant that determine the width of the control area.

While the Shewhart charts are straightforward and interpretable, they can be sensitive to outliers. To address this, MAD – based charts are presented in order to measure deviations from the median rather than the mean, reducing the influence of extreme residuals. Following [21], the control limits are defined as:

$$UCL_{MAD} = median\ (r) + tMAD(r)$$
$$LCL_{MAD} = median\ (r) - tMAD(r)$$

where r is the respective residuals, $MAD(r) = median(|r_i - median(r)|)$, and t is the control limit constant typically set to 3.

Another method for situations where residuals may deviate from normality or exhibit subtle distributional changes KDE – based charts provide a flexible alternative. According to [18], the KDE approximates the probability density function of a residual r as

$$f(r) = \frac{1}{vh}\sum_{i=1}^{v} K\left(\frac{r - r_i}{h}\right)$$

where $K(r) = \exp\left(-\frac{r^2}{2}\right)/\sqrt{2\pi}$ and $h = \left(\frac{4\sigma^5}{3v}\right)^{0.2}$. The cumulative distribution function of residuals estimated from the probability density function is $F(r) = \int_{-\infty}^{r} f(t)dt$ and control limits are determined based on the estimated tails:

$$UCL_{KDE} = F_r^{-1}\left(1 - \frac{\alpha}{2}\right)$$
$$LCL_{KDE} = F_r^{-1}\left(\frac{\alpha}{2}\right)$$

where F_r^{-1} is the inverse cumulative distribution function of the kernel density estimated residual distribution, and $\alpha/2$ is the chosen significance level. KDE-based charts

are particularly effective at detecting small shifts or distributional changes that may be missed by parametric charts.

4 Monitoring Semiconductor Manufacturing Data

4.1 *SECOM* Dataset and Data Preparation

The *SECOM [22]* dataset is publicly available and can be accessed via the UCI machine learning repository. The dataset contains 1,567 samples, each labeled with a binary quality outcome. The response takes the value -1 for a pass and 1 for a fail, following the convention used in the source. Missing or unavailable sensor measurements are denoted by `NaN`**.** These properties, high dimensionality, noisy signals, and missing values, make the *SECOM* dataset a representative dataset for evaluating anomaly – detection and process monitoring methodologies in semiconductor manufacturing.

4.2 Monitoring Procedure

The analysis of the *SECOM* dataset began with preprocessing to ensure data quality and remove redundant or uninformative features. Columns with near-zero variance or a high proportion of missing values, specifically those with more than 50% missing entries, were excluded. It resulted in a feature matrix X with 318 variables and a binary response vector Y. Values of -1 in the response, indicating a pass, were recoded as 0. The dataset was then evenly split into training and test sets, with the first half used to construct baseline models and define control limits. The remaining half was reserved for evaluating the performance of the monitoring methods. To ensure comparability across features and improve numerical stability, both training and test feature matrices were standardized by centering and scaling each variable.

The condition number of the cleaned data, calculated as $CN = \frac{\lambda_{max}(X'X)}{\lambda_{min}(X'X)} = 2.12806 \times 10^{17}$, where λ_{max} and λ_{min} are the highest and lowest eigenvalues of $X'X$, indicates extremely high collinearity among the predictors.

The stepwise procedures for the PCA-based and ridge-penalized approaches for monitoring are presented below. Both were repeated separately using deviance residuals and RQRs.

Stepwise Procedure for PCA-Based Logistic Regression

1. The training set was standardized by centering and scaling each feature.
2. PCA was performed on the standardized training set. The number of principal components was selected to capture 90% of the total variance, resulting in a model that included 112 components.
3. A logistic regression model was fitted through the IRLS algorithm to obtain the MLE of the coefficients.
4. Residuals were derived from the fitted model.
5. Extreme residuals were removed using the interquartile range (IQR) rule to obtain a clean set of residuals.

6. Control limits for the residual-based monitoring charts were determined using the formulas provided in Section 3.
7. The test set was standardized using the mean and standard deviation of the training set to maintain consistency with the training set.
8. The standardized test set was projected into the PCA space using the previously computed loadings from the training set.
9. Fitted probabilities for the test observations were obtained.
10. Residuals were calculated for the test observations to evaluate model performance.
11. Observations exceeding the established control limits were flagged as potential deviations. Both the total number of flagged observations and the first occurrence of a signal (run length) were recorded for further analysis.

Stepwise Procedure for Ridge-Penalized Logistic Regression

1. The training set was standardized by centering and scaling each feature.
2. Ridge estimation was applied, with the ridge penalty parameter fixed at $k = 0.5$.
3. The model was fitted using the cyclic coordinate descent algorithm.
4. Residuals were derived from the fitted model.
5. Extreme residuals were removed according to the interquartile range (IQR) rule to obtain a cleaned set of residuals.
6. Control limits for the Shewhart, MAD-based, and KDE-based charts were determined from the cleaned residuals.
7. The test set was standardized using the mean and standard deviation of the training set.
8. Fitted probabilities for the test set were obtained from the ridge-penalized logistic regression model using the training set parameters.
9. Residuals were calculated for the test observations.
10. Observations exceeding the control limits were flagged as potential deviations, and both the total number of flagged observations and the first occurrence of a signal (run length) were recorded.

For the Shewhart and MAD-based charts, the control limit constants l and t are set to 3, whereas for the KDE-based charts, the significance level α is set to 0.05.

4.3 Monitoring Results

For each method, the lower and upper control limits are reported in Table 1.

Table 2 presents the monitoring results, including the run length until the first out-of-control signal and the total number of flagged observations.

Across both Ridge and PCA models, the KDE-based charts consistently produced narrower control limits compared to the traditional Shewhart and MAD-based charts. As a result, the KDE-based approach detected more residuals as out-of-control and generally exhibited shorter run lengths, indicating higher sensitivity to shifts or anomalies in the residual distribution.

Table 1. Control limits.

Ridge-Penalized approach				
		Shewhart chart	MAD-based chart	KDE-based chart
Deviance residual	LCL	-0.6062	-0.5800	-0.5613
	UCL	-0.1658	-0.1735	-0.2587
Randomized quantile residual	LCL	-2.8335	-2.9729	-1.9144
	UCL	2.7707	2.8589	1.7776
PCA-based approach				
		Shewhart chart	MAD-based chart	KDE-based chart
Deviance residual	LCL	-1.6453	-1.6107	-1.0060
	UCL	-0.6419	-0.6604	-0.8070
Randomized quantile residual	LCL	-2.6680	-2.2617	-2.1336
	UCL	1.2495	1.2617	0.5432

The MAD-based charts performed similarly to traditional Shewhart charts for deviance residuals, providing slightly more robust limits for heavy-tailed or skewed residuals. However, for randomized quantile residuals, KDE-based charts identified substantially more outliers than either Shewhart or MAD-based charts, highlighting their advantage in capturing subtle distributional changes that may be missed by parametric or median-based approaches.

Table 2. Monitoring results.

Ridge-Penalized approach				
		Shewhart chart	MAD-based chart	KDE-based chart
Deviance residual	Run length	247	247	36
	Flagged	2	2	32
Randomized quantile residual	Run length	13	7	7
	Flagged	78	100	117
PCA-based approach				
		Shewhart chart	MAD-based chart	KDE-based chart
Deviance residual	Run length	13	13	7
	Flagged	82	97	160
Randomized quantile residual	Run length	13	13	13
	Flagged	16	20	57

Overall, the results demonstrate that KDE-based control charts offer a flexible, non-parametric alternative for monitoring residuals, capable of detecting both location and distributional shifts more effectively than traditional and robust parametric methods.

5 Conclusion

In this study, we investigated residual-based monitoring approaches for semiconductor manufacturing using both Ridge-penalized and PCA-based logistic regression models. We evaluated traditional Shewhart charts, MAD-based charts, and nonparametric KDE-based charts using deviance residuals and randomized quantile residuals.

Our results show that KDE-based charts consistently provide narrower control limits and detect more out-of-control observations than Shewhart and MAD-based charts. This elevated sensitivity enables the detection of subtle shifts and distributional anomalies in residuals that parametric or median-based methods may overlook. MAD-based charts offer increased robustness for heavy-tailed or skewed residuals, particularly for deviance residuals, but their performance is generally comparable to Shewhart charts. These findings highlight KDE-based residual monitoring as an effective tool for process surveillance in complex industrial settings characterized by high-dimensional, correlated features and potentially non-normal residuals. The approach can support more timely and accurate identification of deviations, improving overall process quality in semiconductor manufacturing.

Disclosure of Interests. The author has no competing interests to declare that are relevant to the content of this article.

References

1. McCullagh, P.: Generalized linear models. 2nd edn. Routledge, London (2019)
2. Marx, B.D., Smith, E.P.: Principal component estimation for generalized linear regression. Biometrika 77(1), 23–31 (1990)
3. Segerstedt, B.: On ordinary ridge regression in generalized linear models. Communications in Statistics: Theory and Methods 21(8), 2227–2246 (1992)
4. Schaefer, R., Roi, L., Wolfe, R.: A ridge logistic estimator. Communications in Statistics: Theory and Methods 13(1), 99–113 (1984)
5. Lee, A.H., Silvapulle, M.J.: Ridge estimation in logistic regression. Communications in Statistics: Simulation and Computation 17(4), 1231–1257 (1988)
6. Le Cessie, S., Van Houwelingen, J.C.: Ridge estimators in logistic regression. Journal of the Royal Statistical Society, Series C (Applied Statistics) 41(1), 191–201 (1992)
7. Özkale, M.R., Lemeshow, S., Sturdivant, R.: Logistic regression diagnostics in ridge regression. Computational Statistics 33, 563–593 (2018)
8. Hawkins, D.M.: Multivariate quality control using regression adjusted variables. Technometrics 33, 61–75 (1991)
9. Hawkins, D.M.: Regression adjustment for variables in multivariate quality control. Journal of Quality Technology 25, 170–182 (1993)
10. Wade, M.R., Woodall, W.H.: A review and analysis of cause selecting control chart. Journal of Quality Technology 25, 161–169 (1993)
11. Loredo, E.N., Jearkpaporn, D., Borror, C.M.: Model based control chart for autoregressive and correlated data. Quality and Reliability Engineering International 18, 489–496 (2002)

12. Skinner, K.R., Montgomery, D.C., Runger, G.C.: Process monitoring for multiple count data using a generalized linear model-based control chart. International Journal of Production Research 41, 1167–1180 (2003)
13. Jearkpaporn, D., Montgomery, D.C., Runger, G.C., Borror, C.M.: Process monitoring for correlated gamma distributed data using generalized linear model-based control charts. Quality and Reliability Engineering International 19, 477–491 (2003)
14. Agresti, A., Amiri, A., Niaki, S.T.A.: A new link function in generalized linear model-based control charts to improve monitoring of two-stage processes with Poisson response. International Journal of Advanced Manufacturing Technology 72, 1243–1256 (2014)
15. Pierce, D.A., Schafer, D.W.: Residuals in generalized linear models. Journal of the American Statistical Association 81(396), 977–986 (1986)
16. Feng, C., Sadeghpour, A., Li, L.: Randomized quantile residuals: An omnibus model diagnostic tool with unified reference distribution. arXiv preprint arXiv:1708.08527 (2017)
17. Abu-Shawiesh, M.O.A.: A simple robust control chart based on MAD. Journal of Mathematics and Statistics 8(1), 37–41 (2008)
18. Lee, W.J., Mendis, G.P., Triebe, M.J., Sutherland, J.W.: Monitoring of a machining process using kernel principal component analysis and kernel density estimation. Journal of Intelligent Manufacturing 31, 1175–1189 (2020)
19. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning. 2nd edn. Springer, New York (2009)
20. Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., Qian, J., Yang, J.: glmnet: Lasso and elastic-net regularized generalized linear models (R package version 4.2.2) [Computer software]. CRAN. https://glmnet.stanford.edu (2025)
21. Adekeye, K.S.: Process capability indices based on median absolute deviation. International Journal of Applied Science and Technology 3(4) (2013)
22. McCann, M., Johnston, A.: Dataset: SECOM. UC Irvine Machine Learning Repository. https://archive.ics.uci.edu/dataset/179/secom