Chapter 8

Open Questions and Conclusions

8.1 Open Questions in High Breakdown Estimation

There are still some unanswered questions related to high breakdown estimation methods for multivariate control charts. For example, because the asymptotic distribution of the $T^2_{mcd,i}$ and $T^2_{mve,i}$ statistics is χ^2_p, it may be useful to study the use of approximate control limits which are much simpler to obtain than those obtained via simulation. This type of study was performed for a T^2 statistic with a variance-covariance matrix based on successive differences by Williams et al. (2006b) to compare the probability of signal using a simulated control limit versus the asymptotic χ^2_p limit. We believe that it is likely that large sample sizes are needed for a χ^2_p approximation to be sufficiently accurate for T^2 statistics based on high breakdown estimation methods.

We have only considered here high breakdown estimation methods that are robust in the sense that they are resistant to a specific type of outlier. The results here apply only to a single cluster of outliers in the same direction. We have not considered clusters of outliers or outliers in different directions. In addition, it is not clear if these high breakdown estimation methods are robust to other departures from the specified assumptions. For example, it is

not clear if the superiority of high breakdown methods is maintained when the data no longer follow a multivariate normal distribution. A tool that could be useful in detecting clusters of outliers even when the data are not normally distributed is cluster analysis. This well known tool in multivariate analysis would be useful in separating clusters of data into their respective groups and thus may be a good Phase I multivariate quality control technique. It apparently has not been applied to a quality control setting.

The out-of-control situation usually just considers changes in the mean vector. An open issue not studied much is what happens when there are changes in the variance-covariance matrix as studied by Levinson, Holmes, and Mergen (2002) and Khoo and Quah (2003, 2004). A robust version of the statistic in Levinson, Holmes, and Mergen (2002) was proposed by Vargas (2005).

8.2 Open Questions in Profile Monitoring with a LMM

In our simulation studies and comparisons we restricted the investigation to certain types of situations and data scenarios. There are a large number of variations that could be considered for the simulations shown here. We discuss briefly some of these variations here.

8.2.1 Uncorrelated Errors and Random Effects

The studies presented within this proposal have shown some situations where the LMM approach is preferred over the LS approach, namely when the data are unbalanced, n is small, or there are data missing at random. We assumed that the data have random effects and considered both correlated and uncorrelated errors but we have not considered the case when the data are correlated with no random effects, that is, where $\mathbf{y}_i \sim MN(\mathbf{X}_i\boldsymbol{\beta}, \mathbf{R}_i)$ and

$\mathbf{R}_i$ follows some non-diagonal structure.

8.2.2 Random and Fixed Effects

Current simulation results are for the random coefficients model (i.e. $\mathbf{Z}_i = \mathbf{X}_i$). We have not yet considered models where not all the coefficients are random, where the $\mathbf{Z}_i$ matrix is a subset of the $\mathbf{X}_i$ matrix. This would include for example, a model where linear profiles fit with a simple linear regression model had similar slopes but different intercepts. We believe that our conclusions hold where not all of the effects are random.

8.2.3 Alternative Error Structures

Our approach has been to model the errors using an AR(1) structure that only depends on the value of σ^2 and ρ. There are many other error structures that could be utilized that depend on a small number of parameters, for example, the CS structure. We believe that regardless of the error structure used, our conclusions would still hold.

8.2.4 Correlated Random Effects

The $\mathbf{D}$ matrix used is a diagonal matrix which means that the random effects are uncorrelated with each other. This assumption is not always realistic in practice. An alternative assumption is to assume that the $\mathbf{D}$ matrix has no specific structure so that there is a non-zero covariance for each pair of random effects. This assumption can cause computational difficulties if there are a large number of effects, but has not been considered. When there are a large number of random effects, they can be modeled by a structure that depends on a smaller number of parameters (such as an AR or CS), just as is often done for the structure

of the errors.

8.2.5 Multiple Linear Regression

We have extended profile monitoring to the application of mixed models based on one regressor. LMM's can be used for the multiple linear regression case where $p > 2$. This increase in the dimension of the vectors that make up the T^2 statistic could be due to higher order coefficients for a single covariate or multiple covariates. It remains to be seen if our conclusions will hold for higher order models.

8.2.6 Outliers

The power studies on the out-of-control performance of the T^2 statistic have been performed for step changes in the mean vector where the method based on the $T^2_{2,i}$ statistic is superior. We have not considered power studies for outliers (at the profile level or within the profile level). It seems clear that the high breakdown estimation methods of Chapter 3 will be superior and of great benefit for this situation. We believe that the robust T^2 statistics obtained via the LMM approach will be equivalent or superior to those obtained via the LS approach. It should be noted that even if the eblups sum to zero, that $T^2_{mve,i}$ and $T^2_{mcd,i}$ values will likely depend on both the estimated fixed effects and predicted random effects, not just on the random effects as was the case for the $T^2_{1,i}$ and $T^2_{2,i}$ values.

8.2.7 Missing Data

Our studies on missing data assume that data are missing at random. However, for repeated measures data, it is sometimes reasonable to presume that the missing observations depend

on a variable that is not observed. For example, if the profiles were to represent human subjects who are measured at repeated time intervals, dropout can occur because subjects are no longer interested in participating. In some cases, the dropout can be the result of an ineffective treatment and the missing observations no longer occur at random. It is much more difficult to model data that are not missing at random.

8.3 Open Questions in Profile Monitoring with NLM Models

All of the questions posed in the previous section for the LMM can similarly be addressed in the NLM model. For example, we limited ourselves to the balanced, equally spaced data scenario but did not consider unbalanced data or data with missing observations. Because of the superiority of the NLM approach for balanced, equally spaced data, we believe that it will retain or even increase its advantage for unbalanced or missing data scenarios. We have not considered a T^2 statistic based on high breakdown estimators that are excellent at detecting multiple outliers as discussed in Chapter 3.

In addition, further research is needed to determine an appropriate test of lack of fit and test of homogeneity of variance when no replicates are available. It would also be useful to determine what would be remedies when there is lack of fit and/or heterogeneity of variance. It is likely to be extremely difficult to obtain estimates in a NLM model with heterogeneous variances.

We believe it is often the case in profile monitoring that a single value is obtained at each of the locations along the profile. Profile monitoring depends heavily on the assumption that the fitted profile fits the data well and it is crucial to check that assumption to ensure

confidence in the later steps of the analysis scheme.

8.4 Other Open Questions

As noted in Section 4.8, the SS residuals can be used to detect outlying observations. Our focus has been on detecting outlying profiles. If it is believed that there are outlying observations within a profile, then it would be useful to consider a robust regression method to fitting the profiles. These methods include least median squares (LMS) regression and least trimmed squares (LTS) regression as discussed in great detail in Rousseeuw and Leroy (1987). Extension of LMS to a nonlinear model was done by Vankeerberghen et al. (1995), but extension of LMS and LTS to handle mixed model data with correlated errors does not appear to have been studied in the literature.

Because our focus has been on Phase I applications, we have not considered performance of using high breakdown estimators or mixed models for Phase II applications. We do not believe that high breakdown methods will be useful for Phase II applications because they only use a portion of all the data and thus are statistically inefficient.

For Phase II monitoring of profiles, it is not clear if the mixed model approach will be applicable. A mixed model works well when there are multiple profiles but it appears that fitting a mixed model to a single profile as would be done in Phase II applications would not give any advantage over simple linear or nonlinear model applicable to a single profile.

In addition, we have only considered a simple situation (step changes) of introducing out-of-control data to profile data. There are many ways that out-of-control data can be introduced. We have only considered changes in the mean vector and have not considered changes in the variance-covariance matrix of the data generated.

The results obtained here assume that the parametric model of choice is the correct model and that it is appropriate for the data at hand. An important issue to be considered is the appropriateness of the proposed methods when the model has not been correctly specified. This model misspecification can have a negative impact on our results. Model misspecification can occur when the data and random effects do not follow multivariate normal distributions or when the number of fixed and random effects have been incorrectly specified.

Our proposed method can be a better way of reducing the profile data to a smaller set of values that characterize the profile, but there are a variety of ways the values can be used. For example, Mahmoud and Woodall (2004) proposed an F test on the obtained coefficients from separate simple linear regression models that was found to be superior to some earlier methods for analyzing linear profiles in Phase I. We believe that the method of Mahmoud and Woodall (2004) could be improved by using the eblups from the LMM rather than by using the estimated coefficients from the separate simple linear regression models. This would ensure that the method works well if the data are unbalanced or missing.

An important extension of this work would be monitoring of profiles via other parametric methods such as generalized linear mixed models (GLMM), or to semi-parametric or non-parametric methods as mentioned by Woodall et al. (2004). These methods undoubtedly will allow for more complicated profiles to be fit and monitored like those shown in Zhou, Sun, and Shi (2006).

8.5 Conclusions

It is important for Phase I multivariate control charts to be based on a high breakdown estimator in order to ensure that outliers are detected and that, as a result, the Phase II control limits will be useful. Both the MVE and MCD estimators are effective in detecting multiple outliers, but each is more advantageous for certain combinations of sample size and the proportion of outliers present. The MVE estimator is preferred for smaller sample sizes and a smaller percentage of outliers while the MCD is preferred for larger sample sizes and/or large percentages of outliers. The simulations and generated control limits presented give useful guidelines about the situations for which each high breakdown approach is most appropriate.

We proposed to monitor linear profiles that have been fit by a LMM to account for random effects and correlated errors. We believe that one can safely ignore the correlation structure if the data are balanced or if we have a larger number of observations. The mixed model approach is most beneficial when the data are unbalanced, the number of observations per profile is small, or there are missing data.

For nonlinear profiles, we proposed to fit them with a NLM model to account for random effects and correlated errors. We found that an approach that uses the separate NL regression models to determine needed random effects works well in setting up the analysis of the profiles with a NLM model. This proposed method uses an easy-to-calculate control limit and thus does not require extensive simulation to obtain the correct control limit as the approach of Williams et al. (2006a). We also found that we can safely ignore the correlation of the errors and concentrate our effort in modeling the random effects. Modeling the random effects allows us to use a Phase I control limit that does not need to be obtained via simulation as would be needed if we only obtained the estimates from separate NL regression models.

Our conclusions for both linear and nonlinear profiles agree with the results of Staudhammer et al. (2005) who modeled profiles in lumber manufacturing using mixed models. They modeled the autocorrelated errors with time series models and concluded that for SPC applications, ignoring the autocorrelation does not make much difference. This is true even though their profile model was rather complicated and there is an obvious autocorrelation in the errors. The number of observations that they have per profile is much larger than the sample sizes considered here ($n > 2000$).

Profile monitoring is a very fruitful area of research. We believe that there is a tremendous reservoir of applications where this methodology can be utilized. Profile monitoring is a tool of the future to match the technology of today that will realize more widespread acceptance as it becomes more readily accessible. We encourage further research and applications of profile monitoring.