Multivariate Bounded Process Adjustment Schemes

Nirmal Govind
Enrique del Castillo

Department of Industrial and Manufacturing Engineering
The Pennsylvania State University
University Park, PA 16802

George Runger
Department of Industrial Engineering
Arizona State University
Tempe, AZ 85287

Mani Janakiram
Intel Corporation
Chandler, AZ 85226

Abstract

We investigate multivariate process adjustment policies in the presence of fixed adjustment costs. A fixed adjustment cost leads to a trade-off between adjusting a process continuously and allowing the process to drift until a bound is violated. The problem of determining the best bounded adjustment policy and the optimal bounded adjustment parameter is addressed. Although the univariate bounded adjustment problem has received interest in the past, there has been little focus on the multivariate case in the literature. It has been shown that in the univariate case, the shape of the optimal adjustment policy is a *dead band.* Determining the shape of the optimal *dead region* and the parameter that defines this region in the multivariate case is a difficult problem. A State-Space model is formulated for the multivariate bounded process adjustment problem and a Kalman Filter based controller is used. With the help of simulation, the optimal bounded adjustment parameter is computed for two specific dead subspace shapes of practical applicability. An investigation of the performance of the simulation-optimization approach is included as the dimensionality of the observations increases. Validation of this approach with the only two analytic results available for deadbands (univariate and bivariate) confirms the accuracy of the optimal solutions found. An illustration in a semiconductor manufacturing process is presented and MATLAB code that implements the methods is developed and made available.

KEYWORDS: Multivariate Process Control, Bounded Adjustment, Adjustment Cost, Optimal Adjustment Policy, Drug dosage management.

1 Introduction

Statistical quality control deals with the application of statistical techniques to monitor and control processes to ensure that the output obtained meets certain preset criteria. If the process does not perform as desired, adjustments are made to the controllable inputs to bring the process to a state of statistical control. The monitoring aspect of quality control is often termed Statistical Process Control (SPC) and the adjustment aspect is more commonly known as Engineering Process Control (EPC) or Automatic Process Control (APC). Economic considerations have long been part of the

SPC/EPC methodology where the objective is to control the process such that the costs incurred in operating the process are minimized.

Usually, a cost is associated with the quality characteristic(s) of interest being off-target. This cost is often assumed to be quadratic and proportional to the squared deviation from target. In certain situations, in addition to the off-target cost, there exists a fixed cost for adjusting the process when it is off-target. This is referred to as a *fixed adjustment cost*. For example, overhauling a running machine may result in a fixed cost irrespective of the condition of the machine. The presence of this fixed cost affects the type of control policy that is optimal from an economic point of view.

In classical control problems, an adjustment is made to the process at each discrete time point when the process is off-target. With the fixed adjustment cost, making an adjustment at each of these time points may not be the optimal control policy. Instead, an adjustment is made when some criterion is met. For example, if there is a single quality characteristic of interest and a single controllable input, the criterion may specify that an adjustment be made to the controllable input when the response (quality characteristic) exceeds a certain pre-defined bound. This is more commonly referred to as the *univariate bounded adjustment* problem. Below, we review the previous work on this problem. In this article, our objective is to investigate similar bounded adjustment problems in the multivariate setting, *i.e.* when there are multiple quality characteristics of interest and multiple inputs that can be used to control the process.

The univariate bounded adjustment problem was studied as early as 1963 by Box and Jenkins. In their paper, Box and Jenkins [2] discuss the "machine tool control problem" where a machine is mass producing components and a quality characteristic is measured periodically. A fixed cost of making an adjustment to the machine was assumed. They showed that the optimal control policy that minimizes the costs of being off-target and making adjustments has the form of a *dead band*, which is a set of symmetric bound lines around the desired target. The optimal policy is to adjust if the one-step-ahead minimum mean squared error (MMSE) forecast of the quality characteristic does not fall in the dead band.

Crowder [5] considers the univariate bounded adjustment problem for the finite horizon case. A dynamic programming approach is used to find the optimal adjustment strategy. It is shown that in the finite horizon case, the optimal policy has time-varying limits. Box and Jenkins used a renewal process argument to derive the optimal adjustment policy for the machine tool problem. Crowder shows that the limits increase as end of the process nears and hence it is less likely that an adjustment will be made near the end of run. The short-run limits are shown to be dependent on the length of the production run and the use of infinite-horizon limits for a short-run problem is shown to increase the total cost.

Box and Kramer [1] extend the methodology to the case where observations are taken at specified sampling intervals and there is a cost of taking an observation. Jensen and Vardeman [11] consider the finite-horizon model in the presence of adjustment errors and deterministic drift in addition to a fixed adjustment cost. They show that the optimal policy in this case also is a deadband of upper and lower action limits that depend on the parameters that define the process. In addition, they show that even with zero adjustment cost, the optimal policy is a deadband in the presence of adjustment errors. Luceño et al. [12] present a numerical method that minimizes the overall expected cost of taking an observation, making and adjustment and being off-target. This method computes the optimal sampling interval, the bound limits and the average adjustment interval in the univariate bounded adjustment problem with fixed adjustment costs and sampling costs.

Del Castillo and Göb [7] consider a bivariate bounded adjustment problem with a quality vector consisting of two responses and an input vector of two controllable inputs. With the help of approximations, they derive the optimal adjustment policy for the bivariate case.

It is important to point out that all of the above methods consider the univariate bounded adjustment problem where there is one response and one input, with the exception of Del Castillo and Göb [7] who consider a bivariate problem. No authors have considered higher dimensional bounded adjustment problems. However, in many practical applications in industry we are forced to deal with multiple quality characteristics and multiple inputs to adjust the process. Healthcare presents another set of interesting applications where multiple responses need to be jointly considered. Doyle III et al. [8] consider a Run-to-run control approach for diabetes management where the maximum and minimum glucose levels of patients need to be controlled by adjusting the insulin timing and quantity. They note that a multivariate control approach should be designed to solve the problem; however, univariate controllers are used for each input since they state that the effects of the inputs are relatively decoupled. Good et al. [9] consider the general drug dosage adjustment problem and show how a run-to-run control approach can be better than existing control algorithms in use via a case study on the control of anticoagulant medication. In such drug therapy management problems where there is a fixed cost of applying a medication or treatment, a bounded adjustment policy that recommends "control" only when certain health indicators (e.g. maximum glucose level in the case of diabetes) exceed certain bounds would be useful in minimizing the overall cost of medication as well as the discomfort suffered by the patient.

It must be noted that an analytical solution to the general multivariate bounded adjustment problem where the *shape* of the optimal region is also found is a difficult problem. In this paper, instead, we focus our attention on two specific the multivariate bounded adjustment policies with given shapes and optimize the parameters given the form of the policy. We do not consider sampling costs in this paper.

The paper is organized as follows. In Section 2, we describe the bounded adjustment problem in more detail and outline the univariate dead band policy and propose two multivariate policies that we investigate. Section 3 discusses the approach that we adopt to solve the multivariate bounded adjustment problem. The simulation approach that is used is described in Section 4. This section includes validation results for our method compared to the only two instances where analytic results have previously been derived (univariate and bivariate). Computational requirements of the simulation-optimization method as the dimensionality increases are also analyzed. To show the practical applicability of the approach presented in this paper, we provide a case study with a multivariate problem in Section 5.

2 Bounded Adjustment Problem

Usually, in standard control problems, an off-target cost is incurred when the quality characteristic deviates from a target, or equivalently, when the deviations from target Y deviate from zero. Often, this cost is assumed to be proportional to the square of the deviation from target. However, in certain processes, a fixed cost may be incurred each time an adjustment is made to the process. This is particularly true in *discrete part industries*, for instance, in metal part, electronics or in semiconductor manufacturing, as pointed out by Box and Luceño [4]. In these industries, to make an adjustment it may be necessary to stop a machine to dismount and re-mount a part or change a tool that wears (e.g., polishing pads in a polishing process in semiconductor manufacturing). This will result in a fixed cost independent of the magnitude of the adjustment. Control problems with fixed adjustment costs have also been discussed in healthcare, in particular, in therapies that require constant monitoring of a patient's health like insulin administration [8].

The fixed cost adds complexity to the control policy that needs to be applied to optimally control the process. In the absence of fixed costs, we are familiar with the notion of continuously adjusting the process (at discrete time intervals) when the process is off-target. In such cases, the optimal control is derived based on some criterion, for example, minimizing the mean squared error (MSE) of the forecast of Y or minimizing the variance of the adjustments made. In the presence of fixed costs, the optimal control policy depends on the costs involved and the procedure to find this optimal policy involves minimizing a loss function that accounts for the cost for being off-target and the fixed adjustment costs.

Below, we briefly describe the univariate bounded adjustment problem and the optimal adjustment policy for this problem. We then outline the multivariate problem and discuss two bounded adjustment policies that we consider in this paper for controlling the process.

2.1 Univariate Dead Band

The univariate bounded adjustment problem was first studied in a seminal paper by Box and Jenkins [2]. They consider the machine tool control problem where a machine mass produces components and some quality characteristic of the component is of interest. Let Y denote deviations of the quality characteristic from its target. A cost is incurred whenever Y is off target (off zero) and this cost is assumed to be proportional to the square of the deviation from target, with the proportionality constant defined by a parameter C_T, a positive constant. In addition, a fixed adjustment cost C_F is sustained each time an adjustment is made. C_F is assumed to be independent of the magnitude of the adjustment made. This fixed cost can be thought of as the price that is to be paid in changing the settings of the controllable inputs or the cost of replacing a part etc.

Box and Jenkins showed that the shape of the optimal bounded adjustment policy in the univariate case was a *dead band*, characterized by a parameter L. They use dynamic programming to show that the optimal control policy is to adjust when the absolute value of the forecast of the quality characteristic exceeds L, *i.e.* when $|\hat{Y}| \geq L$. A renewal reward process argument was used to show that the optimal value of the parameter L is obtained by minimizing the long run expected loss given by:

$$Loss \quad = \quad \frac{C_T E(\sum_t Y_t^2)}{AAI} + \frac{C_F}{AAI} \tag{1}$$

where AAI is the average adjustment interval which is defined as the average number of periods between two adjustments.

Box and Jenkins [2] also show that the optimal L depends on the costs via the ratio $\frac{C_F}{C_T}$ and they provide a short table with optimal parameter values for various values of this ratio. Box and Luceño [4] extend this table further using simulation techniques and employ a numerical method in [12]. We discuss these tables in more detail in a later section.

2.2 Multivariate Bounded Adjustment Policies

We now consider extension of the univariate deadband adjustment case to the case a process is multivariate with both a vector quality characteristic $\mathbf{Y} \in \Re^p$ and a vector of adjustment variables $\mathbf{U} \in \Re^q$. Such situations arise frequently in industries such as semiconductor manufacturing where multiple quality characteristics are measured and multiple input settings need to be adjusted to control the process. We consider the general MIMO (Multiple Input Multiple Output) process adjustment problem in the presence of fixed adjustment costs.

Let us assume that the process is monitored at discrete time points t, $t = 1, 2, \ldots$. Let $\mathbf{Y}_t$ denote the observed vector of deviations from target of the quality characteristics of interest to the process engineer. The process engineer wishes to control the process in order to minimize the deviations from target. This is done by adjusting the vector of controllable factor by amounts denoted by $\mathbf{U} \in \Re^q$, at discrete time intervals. Let us assume that an adjustment $\mathbf{U}_{t-1} \in \Re^q$ is applied after observing $\mathbf{Y}_{t-1}$ and before observing $\mathbf{Y}_t$. We assume a state space formulation, where the adjustments affect the unobserved state $\mathbf{Z}$. We further assume that the process has no dynamics, *i.e.* the effect of a control applied to the state after $t-1$ manifests itself in the observation made at time t.

As noted earlier, due to the presence of fixed costs of adjustment, it may not be optimal, in terms of total cost, to adjust the process at every time period. Hence a bounded multivariate adjustment policy is required to guide the process engineer and determine when an adjustment needs to be made. We note that if the fixed cost $C_F = 0$, the optimal policy in the multivariate case also is to adjust at each time period.

Two issues need to be addressed for the bounded adjustment problem. The first one deals with the type of policy that is used to trigger an adjustment and the second one deals with determining the optimal adjustment parameter for the policy that would minimize the average cost inclusive of adjustment and off-target costs. In the univariate case, the optimal policy was shown by Box and Jenkins to be a dead band where an adjustment is triggered when the forecast of the characteristic exceeds a certain width of the band. In the general multivariate case, the shape of the optimal *dead region* is not clear. Below, we outline two policies that we consider to be reasonable for bounded adjustment. We discuss an approach that finds the optimal parameter for each of these policies in the next section. It must be noted that the multivariate bounded adjustment problem is a considerably difficult one and it may not be possible to find an analytical solution such as the one given by [2] for the univariate case.

The multivariate bounded adjustment policies that we consider are as follows:

1. Quadratic Form (Ellipsoidal)
   Here we consider a policy where an adjustment is made after the observation at time t (and before time $t+1$) if the following condition is met:

$$(\hat{\mathbf{Y}}_{t+1})' B_1 (\hat{\mathbf{Y}}_{t+1}) \quad \geq \quad H \tag{2}$$

   where $\hat{\mathbf{Y}}_{t+1}$ is the minimum mean square error (MMSE) one-step-ahead Kalman filter-base forecast of the deviations from target (different targets for each variable are possible) and B_1 is an arbitrary $p \times p$ positive definite matrix associated with this policy (Policy 1) based on process considerations. The adjustment dead region in this case is an ellipsoid in $\Re^p$.

2. Box Form (Polyhedral)
   The adjustment policy here is a direct extension of the univariate dead band into $p-$dimensional space. The policy triggers an adjustment when the following condition is met:

$$b_i.|\hat{\mathbf{Y}}^{(i)}_{t+1}| \quad \geq \quad \sqrt{H} \qquad \text{for any } i, \ i = 1, \ldots, p \tag{3}$$

   where $\hat{\mathbf{Y}}^{(i)}_{t+1}$ is the i^{th} component of the deviation from target vector at $t+1$ and b_i is a process related parameter corresponding to the i^{th} dimension of the quality vector. The b_i's are specified in the form of a vector $\mathbf{B}_2$ (corresponding to Policy 2) by the user based on process considerations. The above condition implies that an adjustment is made when any component $\mathbf{Y}^{(i)}$ of the quality vector $\mathbf{Y}$ violates the univariate dead band on the one-dimensional subspace specified by i (in other words, the $i-$axis).

Note we use the same notation H to denote either bound regardless of the form of the adjusting policy (ellipsoidal or polyhedral). In the rest of the paper, it will be made clear which of the two policies we are talking about when we refer to H. The ellipsoidal form is probably easier to implement in practice in a non-automated form, as it calls for an adjustment when a scalar quantity violates its bound. It could be implemented with an "adjustment chart" similar to those proposed by Box and Luceño [4] in the univariate case.

In the next section, we describe the multivariate approach to finding the best bounded adjustment policy among the two policies considered above. We discuss the process model, the MMSE predictions using a Kalman filter, and the methodology for finding the optimal bound parameter of each of the two adjustment policies.

3 Multivariate Approach

A State-Space model is formulated for the process and the control action, when the bounded adjustment policy triggers an adjustment, is computed based on a Kalman Filter estimate of the one-step-ahead forecast of $\mathbf{Y}$. We show the equivalence of this approach with a multivariate EWMA controller and compute the corresponding EWMA parameter in a special case.

3.1 Process Model

The State-Space model with control that is used is as follows:

$$\begin{aligned} \mathbf{Y}_t &= A\mathbf{Z}_t + \boldsymbol{\alpha}_t \\ \mathbf{Z}_t &= \mathbf{Z}_{t-1} + \mathbf{U}_{t-1} + \boldsymbol{\beta}_t \end{aligned} \tag{4}$$

where

- $\mathbf{Y}_t$ is a $p \times 1$ vector of observations (deviations from target);
- $\mathbf{Z}_t$ is the $q \times 1$ state vector;
- $\mathbf{U}_t$ is a $q \times 1$ vector of *adjustments* to the controllable factors;
- $\boldsymbol{\alpha}_t$ is a $p \times 1$ vector that corresponds to a multivariate normal white noise series $\sim \mathrm{N}(0, \boldsymbol{\Sigma}_\alpha)$;
- $\boldsymbol{\beta}_t$ is a $q \times 1$ vector that corresponds to a multivariate white noise series $\sim \mathrm{N}(0, \boldsymbol{\Sigma}_\beta)$, (the sequence of $\boldsymbol{\alpha}_t$'s and $\boldsymbol{\beta}_t$'s are assumed to be uncorrelated)
- A is a $p \times q$ matrix that relates the observation to the state.

The State-Space formulation models the internal process dynamics that relate the state variables specified by $\mathbf{Z}$ to the observations specified by $\mathbf{Y}$. The inputs to the process, specified by $\mathbf{U}$ are assumed to affect the state of the process via the state equation. The process dynamics are specified by the measurement equation that relates the state vector to the measurement vector. The multivariate white noise $\boldsymbol{\beta}$ represents the drift in the state $\mathbf{Z}$ and $\boldsymbol{\alpha}$ represents the measurement error in the observation $\mathbf{Y}$. The control $\mathbf{U}_{t-1}$ is applied after the measurement $\mathbf{Y}_t$ is taken and before $\mathbf{Y}_{t+1}$ is observed. It is assumed that given the initial state of the process $\mathbf{Z}_0$, the measurements $\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_t$ up to time t and the inputs $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_t$, estimates of the state and output of the process in the future can be computed. In the next subsection, we briefly describe the use of Kalman Filtering to compute the estimates of the state and output. The control action to be applied depends on the estimates as will be shown below.

3.2 Kalman Filter Based Estimation and Control

The control strategy is based on the estimation of the state using a Kalman filter. A Kalman Filter (KF) is a recursive data processing algorithm that uses the observations of a state-space system to estimate the state, filtering the observational noise. A very large body of literature exists on Kalman filters, see [13] and [16] for an introduction. While other heuristic methods for estimation of the state (and hence, for control), such as an exponentially weighted moving average (EWMA) could be used as well, they would lack general optimality properties. In contrast, the Kalman Filter is known to yield the optimal reconstruction and prediction of the state $\mathbf{Z_t}$ given the observed $\{\mathbf{U_t}\}$ and $\{\mathbf{Y_t}\}$ assuming the model is as in (4). If the errors are normally-distributed, the KF is MMSE optimal, if the errors are not normal the KF still provides the best linear estimator [6]. An interesting relation shown in appendix A and B is that if in (4) the covariance matrices are diagonal and the state and observation vectors are of the same dimension ($p = q$), the KF has *exactly* the form of an EWMA of the observations (with a particular choice of smoothing parameter).

The steps used to predict the state and the observations in the multivariate State-Space approach can be summarized as follows. Starting with an initial value for the state vector, say $\hat{\mathbf{Z}}_0$, observations of the quality vector $\mathbf{Y}$ (dev. from target) are made at discrete time intervals. At each time interval t, the posterior estimate of the state $\hat{\mathbf{Z}}_t$ is computed based on the measurement $\mathbf{Y}_t$ using a KF. This, in turn, is used to construct a one-step-ahead prior estimate $\hat{\mathbf{Z}}^-_{t+1}$ of the state. This prior state estimate is used to compute a one-step-ahead prior estimate $\hat{\mathbf{Y}}^-_{t+1} = A\hat{\mathbf{Z}}^-_{t+1}$. Once the measurement at time $t+1$ is made, the posteriori estimate of the state vector at $t+1$ ($\hat{\mathbf{Z}}_t$) is computed and the procedure repeats itself for the next time period. Meinhold and Singpurwalla [14], Crowder [5], and Del Castillo [6] provide more details regarding the Kalman Filter and its application to statistical quality control.

If $\mathbf{Y}$ is a vector of deviations from target and if there exist only quadratic off-target costs, the optimal control to be applied at every time period t is

$$\mathbf{U}_t = -\hat{\mathbf{Z}}_t \tag{5}$$

where $\hat{\mathbf{Z}}_t$ is the KF estimate of the state vector at t. This is easily justified by examining the state equation since setting $\mathbf{U}_t = -\hat{\mathbf{Z}}_t$ would make the one-step-ahead estimate $\hat{\mathbf{Z}}_{t+1} = \mathbf{0}$ and hence $\hat{\mathbf{Y}}_{t+1} = A\hat{\mathbf{Z}}_t = \mathbf{0}$. When fixed adjustment costs are also present, this controller needs to be modified. We use instead any of the two control decision rules in section 2 and apply (5) *only at the time instances when the rule indicates so, otherwise we let the process run uncontrolled.* Note that the deadband control decision rules use the prediction of the observations, but these are also obtained from the KF prediction of the state, since from the observation equation we have that $\hat{\mathbf{Y}}_{t+1} = A\hat{\mathbf{Z}}_{t+1}$.

We would like to point out that the KF estimate of the observations based on the Kalman Filter is equivalent to an EWMA-based forecast. We show in Appendix A that the following relation holds for the one-step-ahead estimate of $\mathbf{Y}$ obtained using the Kalman Filter:

$$\hat{\mathbf{Y}}_t = AK\mathbf{Y}_t + (\mathbf{I} - AK)\hat{\mathbf{Y}}^-_t$$

where K is the $q \times p$ steady state Kalman gain matrix (refer Appendix A for definition) and $\hat{\mathbf{Y}}^-_t$ is the a priori estimate of the measurement vector prior to the measurement at time t. This can be rewritten as follows:

$$\hat{\mathbf{Y}}_t = \Delta\mathbf{Y}_t + (\mathbf{I} - \Delta)\hat{\mathbf{Y}}^-_t \tag{6}$$

where $\Delta = AK$. Thus the one-step-ahead prediction (without control) is given by a vector EWMA with parameter $\Delta = AK$.

We also consider a special case of the above when the covariance matrices $\mathbf{\Sigma}_\alpha$ and $\mathbf{\Sigma}_\beta$, and A are diagonal and the measurement and state vector are of the same dimension, *i.e.* $p = q$. In Appendix B, we show that it is possible to compute the EWMA parameter analytically.

3.3 Multivariate Loss Function

Extending the univariate loss function in (1) to the multivariate case, the total cost incurred has two components - (1) the cost of being off-target, say L_{T_t}, which is assumed to be quadratic, and (2) the fixed cost of making adjustments, say L_F, where $L_{T_t} = \mathbf{Y}_t' C_T \mathbf{Y}_t$ at each time period t and $L_F = C_F$ each time an adjustment is made. Note that $\mathbf{Y}_t$ denotes a vector of deviations from the target. Then, the overall expected loss for an infinite time-period run can be written as

$$L = E(\sum L_{T_t}) + E(L_F) \tag{7}$$

For computational purposes, we can estimate L per unit time interval with $\hat{L}$ (thus $\hat{L}$ is the overall loss per time unit) in a finite-horizon run of N time periods as follows:

$$\hat{L} = \frac{\sum_t \mathbf{Y}_t' C_T \mathbf{Y}_t}{N} + \frac{C_F \times \hat{N}_A}{N} \tag{8}$$

where $\hat{N}_A$ is the actual number of adjustments made during the run. Note that $E(L_F) = C_F \times E(N_A)$, where N_A is a random variable that represents the number of adjustments made in a run.

For the univariate deadband adjustment problem, Box and Jenkins [2] pointed out that the optimal deadband parameter depends on the ratio $\frac{C_F}{C_T}$ of the fixed cost to the off-target quadratic cost (note that in the univariate case, C_T is a scalar and hence we can divide by this quantity). In the multivariate setting, we use the form of the loss function in (8) to derive an analogous result.

From (8):

$$L = E[\sum(\mathbf{Y}_t' C_T \mathbf{Y}_t)] + C_F \times E(N_A)$$

The expectation of the quadratic form above is obtained by noting that $E(\mathbf{Y}_t' C_T \mathbf{Y}_t) = \boldsymbol{\mu}_{\mathbf{Y}_t}' C_T \boldsymbol{\mu}_{\mathbf{Y}_t} + trace(C_T \mathbf{\Sigma}_{\mathbf{Y}_t}) = trace(C_T \mathbf{\Sigma}_{\mathbf{Y}_t})$ where $\boldsymbol{\mu}_{\mathbf{Y}_t}$ is the mean and $\mathbf{\Sigma}_{\mathbf{Y}_t}$ is the covariance matrix of the vector $\mathbf{Y}_t$. Note that we have chosen $\mathbf{Y}_t$ to represent the deviation from target of the response and hence $\boldsymbol{\mu}_{\mathbf{Y}_t} = \mathbf{0}$. Thus we see that the average loss in the multivariate setting depends on the ratio $\frac{C_F}{trace(C_T \mathbf{\Sigma}_{\mathbf{Y}_t})}$.

We use either of the two control schemes introduced earlier (polyhedral or ellipsoidal) and minimize the loss function $\hat{L}$ with respect to the deadband control parameter H using a simulation approach described next. The simulation approach provides estimates of the two expected values present in L, given that obtaining closed-form expressions for these expectations in the general dimensional case appears to be a formidable problem (see [7]).

4 Simulation-Based Optimization

We use simulation to determine the optimal bound parameter for a given policy and process. The simulation was implemented in Matlab and takes input parameters that define the State-Space model in (4).

4.1 Algorithm

The simulation-optimization algorithm is as follows. Given input parameters A, $\mathbf{\Sigma}_\alpha$, $\mathbf{\Sigma}_\beta$, C_F and C_T, the simulation tries to find the best policy among the two policies described in 2.2 and the value of H that minimizes the overall cost of being off-target and of making adjustments. This is done by optimizing the loss function in (7) with respect to the bounded adjustment parameter H. It is assumed that the model parameters A, $\mathbf{\Sigma}_\alpha$, $\mathbf{\Sigma}_\beta$ have been estimated from past process data (maximum likelihood methods for this task are described in Shumway and Stoffer [15] ch. 4). Designing a multivariate bounded chart without the cost parameters C_F and C_T in case they are hard to estimate is described in section 5.2 below.

A Matlab function simulates the State-Space model for a large number of time periods, say N. At each time period, the state and measurement vectors for the next time period are estimated using the Kalman Filter, the condition specified by the bounded adjustment policy is checked and an adjustment, given by (5), is made if the condition is violated. The Kalman Filter estimation is implemented using Matlab's `kalman` function that is part of the Control System toolbox.

The optimization routine consists of a search routine that searches over a one-dimensional grid for the optimal value of H. The routine is programmed to stop after a pre-defined tolerance condition on the current value of the cost function. A pre-specified *required percentage difference* on the cost can be specified so that the algorithm will stop the line search if the percentage (p.c.) difference between the current cost function and the minimum cost so far exceeds this required p.c. difference.

The computational time to find the optimum bound H^* is not affected too strongly by the dimensionality of the control problem: regardless of the dimensionality, the optimization is univariate, as the control policies have only one parameter, the bound H^*. For a fixed number of periods N, the time to simulate the process depends on the multivariate random generator algorithm used. Also, the computations in the Kalman gain require a $p \times p$ inverse where p is the no. of responses and for large p may become a bottleneck. Table 1 shows the results of running our implementation of the simulation-based line search optimization over H using Matlab on a 4GHz 4-core iMac. The total time to obtain H^* is shown in seconds and scaled by using the time for $p = 3$ as the unit. Since the optimization is stopped when the percentage difference exceeds 5%, it makes more sense to look at the computing time per search point, since as p increases the algorithms searches over more values of H. The values H_{max} indicate what was the largest value tried until the stopping criteria of the algorithm; since the step size was 0.5 there were twice these many search points. The last two columns therefore show the computational time per search point in seconds and scaled relative to the $p = 3$ case. As it can be seen, the increase in total computing time is in good part due to the longer searches needed as p increases, since H^* will typically be higher for p higher. The per point times begin to increase faster than linearly when the inverse operations ($O(p^3)$) become the bottleneck. We point out,however, that this optimization is done *off-line*, only once. The operation of the resulting multivariate bounded controller does not requires major computational effort and can be done on-line regardless the dimensionality of p or q.

The Matlab code `muBoundS` implements the optimization routine to find H^* and is available for download from the Engineering Statistics and Machine Learning Laboratory at the Pennsylvania State University: `http://sites.psu.edu/engineeringstatistics/computer-codes/`. More information on the use of the different options of the optimization routine such as the termination conditions can be found in the README.txt file at the same location.

p	H_{max}	total time	scaled total	search points	time/point	scaled per point
3	11.50	8.75	1.00	23	0.38	1.00
5	13.50	11.02	1.26	27	0.41	1.08
10	18.50	16.06	1.84	37	0.43	1.13
20	23.50	22.58	2.58	47	0.48	1.26
40	37.00	48.58	5.55	74	0.65	1.71
70	62.50	122.03	13.95	125	0.97	2.55
100	90.00	245.09	28.01	180	1.36	3.58
150	136.50	642.90	73.47	273	2.35	6.18

Table 1: Computational effort for finding the optimal bound H^* as a function of p, the dimension of the observation vector. Unscaled times in secs. Parameters used were $p = q$, $A = \mathbf{\Sigma}_\alpha = \mathbf{\Sigma}_\beta = C_T = B_1 = \mathbf{I}_p$, $C_F = 10$, $N = 2500$ (first 500 eliminated), 50 replicates, ellipsoidal policy. The search on H used a grid size equal to 0.5, and stopped when the percentage difference in cost exceeded 5% (see text).

4.2 Validation

Since the optimization approach is based on simulation, it is important to validate its results comparing against analytic ones. There exist analytic expressions for the loss function only in the univariate and bivariate cases, and in this section we compare our approach against these two cases. In the univariate case, our code was validated comparing against the model originally considered by Box and Jenkins. Box and Luceño [4] augment the table in [2] and provide additional optimal bounded adjustment parameters and corresponding AAI's for various values of the cost ratios $\frac{C_F}{C_T}$. Below, we repeat this table, adding columns for the optimal H computed from the standardized bound $\frac{L}{\lambda\sigma_a}$ in [4], and the optimal H and AAI obtained using the State-Space model with Kalman Filter based estimation.

We see from Table 2 on the following page that the simulation code using the State-Space model with Kalman Filter-based estimation was able to validate the AAI and the MSD reported by [4] (using the optimal bound parameter supplied by [4]). The columns of the table are as follows. BL H_{opt} and BL AAI refer to the optimal value of H (computed from their optimal parameter L) and the AAI at this optimal value respectively as given by [4]. The optimal values computed by our simulation code, using the polyhedral scheme, are given by H_{opt} and AAI. The minimum value of the cost function at our H_{opt} is given by $L_{H_{opt}}$.

In order to verify that our simulation code can match the [4] results exactly, we computed the $AAI_{BL\ H_{opt}}$, mean squared deviation $MSD_{BL\ H_{opt}}$ and cost $L_{BL\ H_{opt}}$ using our simulation code at the optimal value BL H_{opt} given by [4]. These were compared with BL AAI and MSD_{BL} provided by [4] and we see that there is a very close match. In addition, we see that there is a good match between our simulation's optimal cost $L_{H_{opt}}$ and the minimum cost computed using BL H_{opt}. Hence we conclude that our simulation code is error-free as far as the univariate case is concerned.

However, we note that the optimal H computed by our simulation optimization differs slightly from the optimal value reported by [4]. We investigated this further with the help of extensive simulation runs. In a simulation experiment with $N = 10,000$, shown in Figure 1 on page 12, we see that the cost function is very flat near the optimum. This leads us to the conclusion that the small difference in optimal values of H between our simulation and the ones provided by [4] is primarily due to the stochastic nature of simulation runs; a different random number seed may very well have resulted in an optimum that is closer to the value reported by these authors. It is also worth mentioning here that our simulation runs used a finite-horizon of $10,000$ time periods and this

R_A	C_F	BL $\frac{L_{opt}}{\lambda\sigma_a}$	BL H_{opt}	H_{opt}	BL AAI	AAI	$AAI_{BL\ H_{opt}}$	$L_{BL\ H_{opt}}$	$L_{H_{opt}}$	MSD_{BL}	$MSD_{BL\ H_{opt}}$
1	0.25	0.93	0.22	0.28	2.6	2.9	2.57	1.13	1.13	1.04	1.04
2	0.5	1.23	0.38	0.4	3.5	3.7	3.54	1.22	1.22	1.08	1.07
3	0.75	1.43	0.51	0.48	4.3	4.1	4.28	1.29	1.28	1.11	1.11
4	1	1.59	0.63	0.64	4.9	5.0	4.97	1.33	1.33	1.14	1.13
5	1.25	1.72	0.74	0.76	5.5	5.6	5.54	1.38	1.38	1.16	1.16
10	2.5	2.17	1.18	1.1	7.8	7.4	7.85	1.57	1.57	1.25	1.25
20	5	2.70	1.82	2.0	11.0	11.8	11.06	1.83	1.83	1.38	1.38
30	7.5	3.05	2.33	2.4	13.5	13.9	13.47	2.04	2.03	1.48	1.49
40	10	3.33	2.77	2.55	15.5	14.6	15.50	2.22	2.21	1.58	1.58
50	12.5	3.55	3.15	2.9	17.3	16.3	17.43	2.37	2.36	1.65	1.64
60	15	3.75	3.52	3.3	19.0	18.2	19.17	2.51	2.49	1.73	1.71
70	17.5	3.92	3.84	3.7	20.6	19.9	20.68	2.62	2.62	1.78	1.77
80	20	4.07	4.14	4.0	21.9	21.5	21.92	2.74	2.73	1.83	1.84
90	22.5	4.21	4.43	4.3	23.2	22.8	23.26	2.87	2.84	1.9	1.89
100	25	4.34	4.71	4.9	24.5	25.6	24.52	2.96	2.94	1.95	1.94
200	50	5.28	6.97	6.7	34.7	33.9	35.00	3.81	3.78	2.38	2.36
300	75	5.91	8.73	9.1	42.4	44.8	42.68	4.45	4.41	2.7	2.68
400	100	6.40	10.24	9.9	49.0	48.5	49.43	5.00	4.94	2.98	2.96
500	125	6.81	11.59	11.2	55	54.5	55.21	5.49	5.42	3.2	3.22
600	150	7.14	12.74	13.1	60	62	60.24	5.91	5.86	3.43	3.41
700	175	7.45	13.88	14.0	65	67	66.24	6.31	6.25	3.63	3.6
800	200	7.72	14.89	14.7	69	70	70.33	6.69	6.62	3.8	3.79
900	225	7.98	15.92	15.5	74	73	74.93	7.04	6.96	4	3.97
1000	250	8.21	16.85	16.4	77	77	77.53	7.39	7.29	4.15	4.17
2000	500	9.88	24.40	21.8	110	102	110	10.09	9.94	5.5	5.47
3000	750	11.02	30.36	27.0	135	127	138	11.93	11.83	6.53	6.51
4000	1000	11.85	35.11	33.0	155	151	158	13.79	13.56	6.63	7.39

Table 2: Validation: comparison with univariate results in [4].

may skew the results a little. However, given the amount of computing time that each optimization run takes (approximately 5 hours on a multi-processor Linux cluster) with this horizon, we argue that a longer run is unwarranted.

Del Castillo and Göb [7] provide results for a bivariate bounded adjustment problem. We validated our Matlab code against the analytical results provided by these authors. They use a bounded adjustment policy that uses a rotated square with half side length Λ as the dead area. Hence our Matlab code was modified to use the rotated square as the dead area instead of the polyhedral or ellipsoidal regions.

The model used for validation purposes was a special case of the model in reference [7] with $\Theta = \mathbf{0}$ in that paper. The resulting process is a bivariate random walk. The parameters of our State-Space approach were adjusted to match the values in [7]. We have $A = \mathbf{I}$, $\Sigma_\beta = \begin{pmatrix} 100 & 10\rho \\ 10\rho & 1 \end{pmatrix}$ and $\alpha_t = 0$ (complete state information).

Table 3 shows the results obtained by [7] and the corresponding results from our Matlab code. In the table, C' represents the relative fixed adjustment cost, ρ is the standardized correlation between the elements of the bivariate white noise series, "C-G AAI" refers to the AAI at the optimum obtained by [7] and "SS AAI" refers to the AAI value obtained by our State-Space based simulation. The simulations were performed for the case where $\rho = 0.6$. The agreement for $\Lambda^* \geq 2$ is very good (for the single case where $\Lambda^* \leq 2$, [7] used certain approximations that put their result and not ours in doubt.)

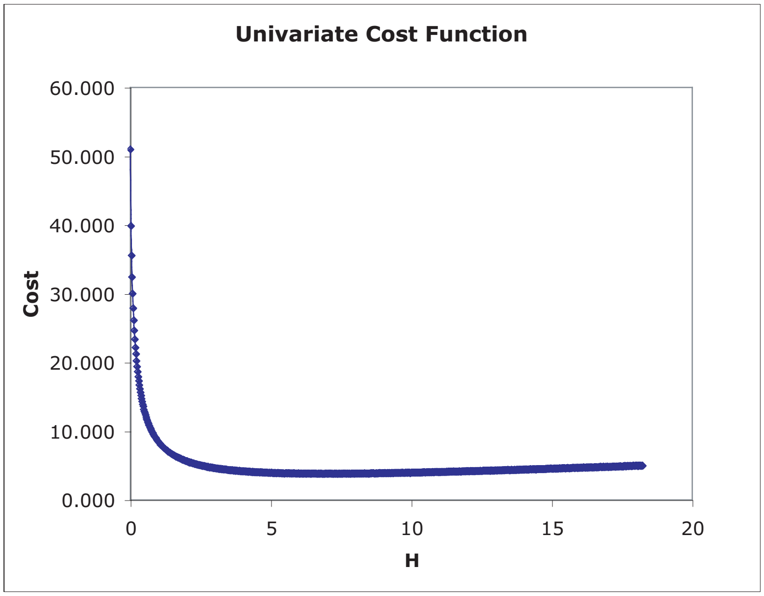

Figure 1: Univariate Cost Function with $N = 10,000$

C'	$\|\rho\|$	Λ^*	C-G AAI	SS AAI	SE
1	.6	1.21	2.11	1.63	0.0012
4	.6	2.73	4.12	4.08	0.0059
7	.6	3.33	5.47	5.44	0.0079
10	.6	3.74	6.51	6.48	0.0111
20	.6	4.64	9.17	9.14	0.0227
50	.6	6.11	14.47	14.45	0.0284
80	.6	7.01	18.30	18.24	0.0378
100	.6	7.47	20.47	20.46	0.0445
400	.6	11.04	41.13	41.42	0.1331
700	.6	12.89	54.64	54.80	0.1846
1000	.6	14.22	65.55	66.55	0.2631

Table 3: Validation: comparisons with the bivariate process results in [7].

5 An illustration in a semiconductor manufacturing processing step

In this section, we demonstrate the approach that we have outlined in the previous sections using a simulated multivariate model based on a Vertical Diffusion Furnace process in the semiconductor manufacturing industry. In this process, material is deposited on a batch of silicon wafers (usually around 125-150 wafers) and the objective is to maintain a uniform thickness of the deposited material on wafers in each of the zones of the furnace. This is done by controlling the temperature and deposition time. The fixed cost of adjustment in such a process is usually the impact on throughput time of making an adjustment. For example, if a certain amount of time is needed to reach the set temperature, this has a direct impact on the throughput of the tool. The process therefore has three responses and two inputs. The objective is to control the process and bring the responses to desired target values by adjusting these two inputs. Note that the parameters chosen below for the simulation case study do not reflect the actual parameters of the modeled process.

We investigate the performance of the ellipsoidal and polyhedral policies as given by (2) and (3) respectively. From a practical standpoint, we try to take into account the fact that the user may not have complete information on all the costs involved but may still wish to control the process using a bounded adjustment policy for other reasons such as stability of the process. Hence we present two cases - the first case is where the cost information is readily available and the user can obtain the optimal adjustment parameter for each policy as shown in 5.1; the second case is where the cost information is not available (or is incomplete) and the user can pick an adjustment policy and adjustment parameter based on the MSD and AAI values as explained in 5.2. The software developed can handle both these cases and we refer the reader to the software documentation for details regarding the implementation.

The parameters of the model that we use in this simulation case study are as follows:

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 2 & 3 \end{bmatrix}, \; \mathbf{\Sigma}_\alpha = \begin{bmatrix} 4 & 4 & 3 \\ 4 & 6 & 5 \\ 3 & 5 & 6 \end{bmatrix}, \; \mathbf{\Sigma}_\beta = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$$

$$C_T = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 2 & 2 \\ 1 & 2 & 4 \end{bmatrix}, \; B_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \; B_2 = \frac{1}{\sqrt{3}} \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \tag{9}$$

Note that the selection of B_2 was made so that each of the three responses have equal weights and $\|B_2\| = 1$. However, this may not be the case always and this parameter can be changed; the current example is only an illustration.

5.1 Bounded Adjustment With Costs

We first consider the problem of determining the minimum cost optimal bound parameter H for various values of C_F. As an illustration of the algorithm's performance for the case of the Ellipsoidal policy (referred to as Policy 1), Figure 2 on the following page depicts the cost along the line search path for various values of H. The minimum for this case ($C_F = 60$) is at 12.1. The series labeled "True Cost" represents the cost obtained from simulation for various values of H. In order to find the minimum, a cubic function (series labeled "Fitted Cost") was fit to these simulated data points. The minimum was obtained by minimizing this cubic function. The equation for the cubic function is also shown in Figure 2 on the next page.

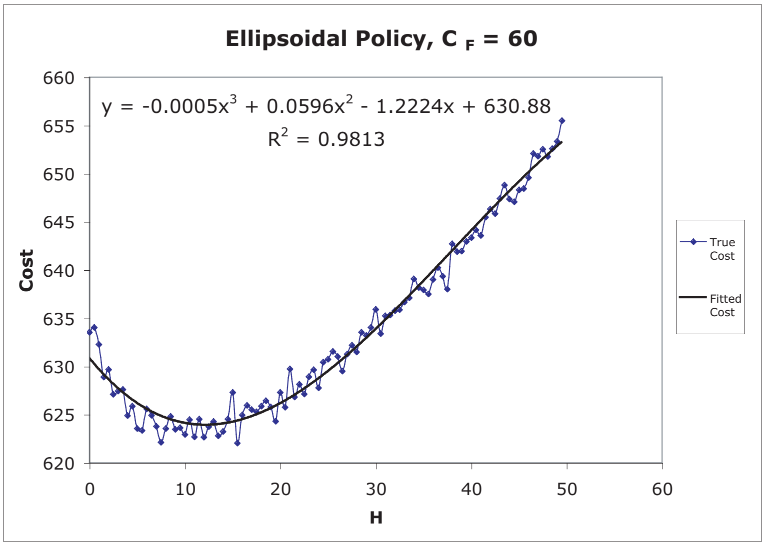

Figure 2: Search path for Ellipsoidal Policy with $C_F = 60$

Table 4 provides the optimal bound parameter H_{opt}, AAI, MSD and minimum cost for various values of C_F while the other parameters are held constant at the values specified in (9). As we can see, the optimal bound parameter H_{opt} increases with C_F, this behavior is the same as in the univariate case and lends itself to the same reasoning that the higher fixed cost of making an adjustment makes it profitable to let the process drift more before an adjustment is made.

It is also interesting to note from Figure 2 that the time it takes until the specified percentage cost difference (5% in this case) is achieved is significantly higher than in the case of the Polyhedral policy that we shall examine below.

C_F	H_{opt}	AAI	MSD	Cost
10	0.9	1.05	91.55	584.3
20	3.7	1.17	91.79	592.9
30	6.4	1.27	92.1	600.9
40	8.0	1.32	92.35	608.8
50	11.3	1.42	93.02	617.5
60	12.1	1.44	93.11	624.3
70	14.5	1.51	93.13	628.7
80	15.8	1.55	93.37	635.4
90	17.2	1.58	94.06	644.9
100	19.7	1.65	94.38	650.4

Table 4: Multivariate Results for Policy 1 (Ellipsoidal)

Table 5 on the following page shows the results from the optimization using the Polyhedral bounded adjustment policy for various values of the fixed cost C_F. In addition to the optimal parameter H_{opt}, the AAI, MSD and minimum cost are also provided. The search path of the line search that was used is illustrated in Figure 3 on the next page.

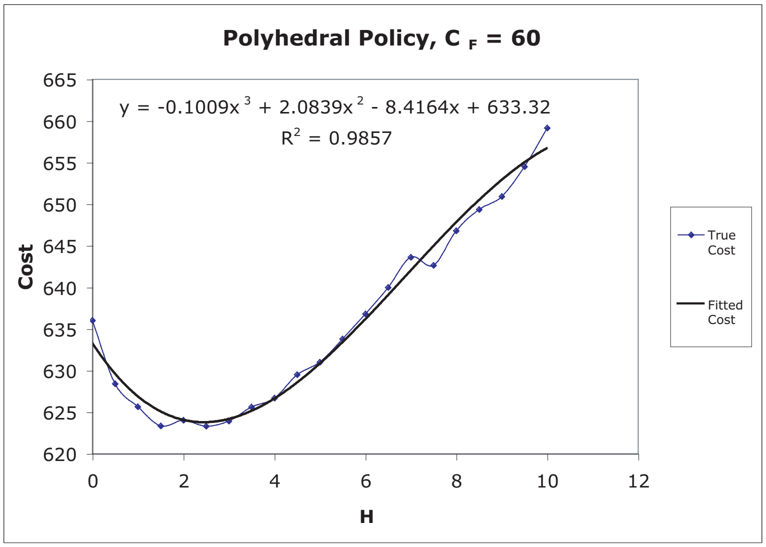

Figure 3: Search path for Polyhedral Policy with $C_F = 60$

A third order polynomial regression function was fit to the simulated data points (labeled as "True Cost"). The minimum cost was obtained by minimizing this cubic function, the optimum value of the bound parameter in this case ($C_F = 60$) is given by $H_{opt} = 2.46$ (from Table 5).

C_F	H_{opt}	AAI	MSD	Cost
10	0	1	91.43	584.1
20	0.8	1.2	91.86	592.9
30	1.35	1.28	92.12	600.94
40	1.7	1.34	92.41	608.92
50	2.25	1.42	92.77	616
60	2.46	1.45	93.07	623.86
70	2.85	1.51	93.42	630.94
80	3.03	1.53	93.5	637
90	3.5	1.6	94.05	644.4
100	3.81	1.64	94.37	650.92

Table 5: Multivariate Results for Policy 2 (Polyhedral)

The behavior of the two adjustment schemes, polyhedral and ellipsoidal, is qualitatively similar: as the adjustment cost C_F decreases, the width of the adjusting bound $H_{opt} \to 0$ implying more frequent adjustments (so AAI $\to 1$), i.e., both control policies appear to converge to a MMSE controller that calls for an adjustment at every point in time t using the KF estimate.

5.2 Bounded Adjustment Without Costs

Consider the case where the user does not have any information on the fixed or off-targets costs involved but is still interested in implementing a multivariate bounded adjustment policy. Often, in industry, process engineers do not wish to adjust the process very often due to concerns that this may make the process unstable. It may also be the case that the fixed cost of adjustment is negligible but a bounded adjustment policy is still preferred due to the same reason.

To account for the case where the cost information is not readily available or the fixed cost is not significant, we make use of the same methodology to obtain results that will aid the process engineer in selecting an appropriate policy based on preferred MSD and AAI values.

Table 6 provides the AAI and MSD with bounded adjustment using the Ellipsoidal policy (Policy 1) for various values of H. From this table, the user can choose the value of H based on the MSD and AAI values that are most acceptable.

H	AAI	MSD
1	1.05	91.6
2	1.1	91.62
3	1.15	91.87
4	1.19	91.92
5	1.22	91.62??
10	1.38	92.35
20	1.65	94.6
30	1.9	96.61
40	2.14	98.85
50	2.37	101.28

H	AAI	MSD
60	2.6	103.62
70	2.83	106.21
80	3.06	108.59
90	3.28	111
100	3.49	113.49
200	5.58	137.61
300	7.57	160.03
400	9.47	181.89
500	11.27	204.07
1000	20.16	307.42

Table 6: Finding a bound limit without costs: AAI and MSD for Policy 1 (ellipsoid)

Table 7 does the same for the case where the bounded adjustment policy used is the Polyhedral policy (Policy 2). Using Table 6 and Table 7, a process engineer can determine which policy is preferable and select the corresponding value of H. Tables that estimate AAI and MSD without cost information are also generated with the aforementioned MATLAB code.

H	AAI	MSD
1	1.22	91.91
2	1.38	92.67
3	1.53	92.75
4	1.66	94.68
5	1.8	95.45
10	2.42	101.44
20	3.57	114.62
30	4.67	127.24
40	5.73	139.18
50	6.74	151.5

H	AAI	MSD
60	7.76	163.02
70	8.78	174.14
80	9.76	185.74
90	10.72	196.45
100	11.63	208.3
200	20.85	315.33
300	29.93	419.75
400	38.43	522.47
500	47.23	622.55
1000	89.03	1121.34

Table 7: Finding a bound limit without costs: AAI and MSD for Policy 2 (polyhedral)

6 Conclusions

The general multivariate bounded process adjustment problem in the presence of a fixed adjustment cost and quadratic off-target cost was studied with the help of a State-Space model and the optimal adjustments have been computed with a Kalman Filter estimator. Two adjustment policies - ellipsoidal (quadratic) and polyhedral (box) - were investigated as possible choices for a dead region within which an adjustment is not made. Simulation-based optimization was used to compute the optimal adjustment parameter for each of these two policies given that closed-form expressions for the expectations in the loss function are not available in the general-dimensional case. A Matlab code for the optimization of these deadband adjustment problems was developed and made available. The method was validated against the only two cases where analytic expressions exist for the expectations in the loss function (a one and a two dimensional process). The methodology applied to a general dimensional problem was illustrated with application to a semiconductor manufacturing process (applications in drug delivery in healthcare were also reviewed in the paper). Methodology was provided for finding the best adjustment deadband bound in the practical case where cost information is not readily available.

This paper developed optimal adjustment strategies for two *given* shapes of the adjustment region that are of practical interest and demonstrated its applicability. Further work is required to determine the shape of the optimal dead region, a difficult problem which is beyond the scope of the current paper.

7 Acknowledgments

This research was supported by NSF GOALI award 0084909. We wish to thank Dr. Russell Barton for helpful suggestions on the paper. In addition, we are grateful to Intel Corporation for hosting one of the authors for a semester (as part of the NSF GOALI plan) while this work was completed. The authors wish to thank an anonymous referee for his/her comments that helped improve our presentation.

References

[1] G. Box and T. Kramer. Statistical process monitoring and feedback adjustment - a discussion. *Technometrics*, 34(3):251–267, August 1992.

[2] G. E. P. Box and G. M. Jenkins. Further contributions to adaptive quality control: Simultaneous estimation of dynamics: Nonzero costs. *Bulletin of the International Statistical Institute*, 34:943–974, 1963.

[3] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis: Forecasting and Control*. Prentice Hall, Englewood Cliffs, NJ, 3 edition, 1994.

[4] G. E. P. Box and A. Luceño. *Statistical Control by Monitoring and Feedback Adjustment*. John Wiley & Sons, New York, NY, 1997.

[5] S. V. Crowder. An spc model for short production runs: Minimizing expected cost. *Technometrics*, 34(1):64–73, February 1992.

[6] E. Del Castillo. *Statistical Process Adjustment for Quality Control*. John Wiley & Sons, Inc., New York, NY, 2002.

[7] E. Del Castillo and R. Göb. A bivariate dead band process adjustment policy. Technical report, Engineering Statistics Laboratory, http://lb350e.ie.psu.edu/publications.htm, IME Department, Pennsylvania State University, 2003.

[8] I. Doyle, F.J., B. Srinivasan, and D. Bonvin. Run-to-run control strategy for diabetes management. *Proceedings of the 23rd Annual EMBS international conference*, pages 3159–3162, 2001.

[9] R. Good, J. Hahn, T. Edison, and J. Qin. Drg dosage adjustment via run-to-run control. *Proceedings of the American Control Conference*, pages 4044–4049, 2002.

[10] J. Hamilton. *TIme Series Analysis.* Princeton University Press, 1994.

[11] K. L. Jensen and S. B. Vardeman. Optimal adjustment in the presence of deterministic process drift and random adjustment error. *Technometrics*, 35(4):376–389, November 1993.

[12] A. Luceño, F. J. González, and J. Puig-Pey. Computing optimal adjustment schemes for the general tool-wear problem. *Journal of Statistical Computation and Simulation*, 54:87–113, 1996.

[13] P. S. Maybeck. *Stochastic models, estimation, and control*, volume 1. Academic Press, 1979.

[14] R. J. Meinhold and N. D. Singpurwalla. Understanding the kalman filter. *The American Statistician*, 37(2):123–127, 1983.

[15] R. Shumway and D. Stoffer. *Time Series Analysis and its Applications.* Springer, 2000.

[16] G. Welch and G. Bishop. An introduction to the kalman filter tr 95-041. Technical report, Department of Computer Science, University of North Carolina at Chapel Hill, March 2002.

Appendix A Equivalence of EWMA to Kalman Filter

The equivalence of the EWMA to a Kalman Filter is derived in this appendix. The steady-state matrices for the usual Kalman Filter estimation problem are:

$$\begin{aligned} K_\infty &= K = PA'(APA' + \Sigma_\alpha)^{-1} \\ P_\infty &= P = P + \Sigma_\beta - PA'(APA' + \Sigma_\alpha)^{-1}AP \end{aligned} \tag{A.1}$$

The steady state above is always reached for arbitrary A and arbitrary positive semidefinite and symmetric Σ_α and Σ_β (see [10] proposition 13.1, p. 390). If $U_{t-1} = 0$,

$$\hat{\mathbf{Z}}_t = \hat{\mathbf{Z}}_t^- + K_t(\mathbf{Y}_t - A\hat{\mathbf{Z}}_t^-)$$

In steady state we have that

$$\hat{\mathbf{Z}}_t = \hat{\mathbf{Z}}_t^- + K(\mathbf{Y}_t - A\hat{\mathbf{Z}}_t^-)$$

$$\hat{\mathbf{Z}}_t = K\mathbf{Y}_t + (\mathbf{I} - KA)\hat{\mathbf{Z}}_t^-$$

and

$$\hat{\mathbf{Y}}_t = A\hat{\mathbf{Z}}_t = AK\mathbf{Y}_t + A(\mathbf{I} - KA)\hat{\mathbf{Z}}_t^-$$

By re-arranging the last term,

$$\hat{Y}_t = AKY_t + (\mathbf{I} - AK)A\hat{Z}_t^-$$

and since $\hat{\mathbf{Y}}_t^- = A\hat{\mathbf{Z}}_t^-$, the above reduces to:

$$\hat{\mathbf{Y}}_t = AK\mathbf{Y}_t + (\mathbf{I} - AK)\hat{\mathbf{Y}}_t^-$$

Hence,

$$\hat{\mathbf{Y}}_t = \Delta\mathbf{Y}_t + (\mathbf{I} - \Delta)\hat{\mathbf{Y}}_t^- \tag{A.2}$$

where $\Delta = AK$.

Therefore, the one-step ahead prediction (without control) is given by a vector EWMA with parameter $\Delta = AK$. $\Box$

Appendix B Special Case: Computing EWMA parameter from State-Space Model

To establish the equivalence of the Kalman Filter and EWMA estimates, we consider a special case with $p = q$, A, P diagonal, $\Sigma_\alpha = \alpha\mathbf{I}$, $\Sigma_\beta = \beta\mathbf{I}$. We first find $\Delta = AK$, the EWMA parameter in (A.2), by solving the steady state Kalman Filter equations (A.1). We then compare the Δ we obtain with the EWMA parameter for the prediction equation obtained using the IMA time series form of the State-Space model. Let $A = D = \{d_{ij} : d_{ij} = 0 \text{ if } i \neq j\},\ i, j = 1, \ldots, q$, a diagonal $q \times q$ square matrix with d_{ii} as the i^{th} diagonal element. Similarly, let $P = \{p_{ij} : p_{ij} = 0 \text{ if } i \neq j\},\ i, j = 1, \ldots, q$, a diagonal $q \times q$ square matrix with p_{ii} as the i^{th} diagonal element. Finally, let $\Sigma_\alpha = \alpha\mathbf{I}$ and $\Sigma_\beta = \beta\mathbf{I}$ be diagonal covariance matrices for the random error terms $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$. Using the Kalman Filter equations, we have from (A.1),

$$P_\infty = P = P + \Sigma_\beta - PA'(APA' + \Sigma_\alpha)^{-1}AP$$

i.e.

$$\Sigma_\beta = PA'(APA' + \Sigma_\alpha)^{-1}AP$$

Substituting the values for A, P, Σ_α, Σ_β and equating elements in the matrices on the two sides, we obtain

$$\beta = \frac{p_{ii}^2 d_{ii}^2}{d_{ii}^2 p_{ii} + \alpha}$$

which reduces to a quadratic equation in p_{ii}. On solving this equation we have

$$p_{ii} = \frac{\beta \pm \sqrt{\beta^2 + 4\frac{\alpha\beta}{d_{ii}^2}}}{2}$$

From (A.1), we have

$$
\begin{aligned}
\Delta = AK &= APA'(APA' + \Sigma_\alpha)^{-1} \\
&= A\Sigma_\beta (AP)^{-1} \\
&= diag(\frac{\beta}{p_{ii}}) \qquad \text{(B.1)}
\end{aligned}
$$

We now obtain the EWMA parameter using an alternative approach. The model in (4) can be written as a vector IMA(1,1) process as follows:

$$(1-B)\mathbf{Y}_t \;=\; (\mathbf{I} - (\mathbf{I} - \Gamma)B)\,\varepsilon_t$$

where matrix Γ is diagonal. The EWMA parameter is given by $\Theta = (\mathbf{I} - \Gamma)$. The EWMA prediction equation for this model is given by ([3], page 151):

$$\hat{\mathbf{Y}}_t \;=\; \Lambda \mathbf{Y}_t + (\mathbf{I} - \Lambda)\hat{\mathbf{Y}}_{t-1}$$

where $\Lambda = \mathbf{I} - \Theta = \Gamma$. The elements γ_{ii} of the diagonal matrix Γ can be computed using the results presented in [3], page 129:

$$\frac{\gamma_{ii}^2}{1-\gamma_{ii}} \;=\; \frac{\beta}{\alpha} a_i' a_i$$

where a_i' is the i^{th} row of the matrix $A = D$ and hence $a_i' a_i = d_{ii}^2$. The above reduces to a quadratic equation in γ_{ii} which can be solved to yield:

$$
\begin{aligned}
\gamma_{ii} &= \frac{\frac{-d_{ii}^2\beta}{\alpha} \pm \sqrt{\frac{d_{ii}^4\beta^2}{\alpha^2} + 4\frac{d_{ii}^2\beta}{\alpha}}}{2} \\
&= \frac{2\beta}{\beta \pm \sqrt{\beta^2 + 4\frac{\alpha\beta}{d_{ii}^2}}} \\
&= \frac{\beta}{p_{ii}}
\end{aligned}
$$

Thus we have

$$\Gamma \;=\; diag(\frac{\beta}{p_{ii}}) = AK = \Delta \qquad \text{(B.2)}$$

Therefore, from (B.1) and (B.2), we see that the EWMA parameter obtained using the two methods is exactly the same. $\square$.