Bayesian Predictive Optimization of Multiple and Profile Response Systems in the Process Industry: a Review and Extensions

Enrique del Castillo*
Dept. of Industrial Engineering and Dept. of Statistics
The Pennsylvania State University, University Park, PA 16802, USA
and
Marco S. Reis
CIEPQPF, Department of Chemical Engineering
University of Coimbra, Coimbra Portugal 3030-790.

June 4, 2020

Abstract

Bayesian statistical methods provide a sound mathematical framework to combine prior knowledge about variables of importance in a process, if available, with the actual data, and has proved useful in several data analytic tasks in the process industry. We present a review and some extensions of bayesian *predictive* methods for process optimization based on experimental design data, an area that is critical in Quality by Design activities and where the bayesian perspective has received limited attention from the Chemometrics and process analytics communities. The goal of the methods is to maximize the probability of conformance of the predicted responses to their specification limits by varying the process operating conditions. Optimization of multiple response systems and of systems where the performance is given by a curve or "profile" are considered, as they are more challenging to model and optimize, yet are increasingly common in practice. We discuss the particular case of Robust Parameter Design, a technique due to G. Taguchi and popular in discrete manufacturing systems, and its implementation within a bayesian optimization framework. The usefulness of the models and methods is illustrated with three real-life chemical process examples. MATLAB code that implements all methods and reproduces all examples is made available.

Keywords: High dimensional response optimization; Quality by Design; Hierarchical Linear Models; Robust Parameter Design.

*Corresponding author, e-mail: exd13@psu.edu

1 Introduction: Bayesian optimization and Robust Parameter Design

The utility of the bayesian paradigm to model uncertainties in complex engineering systems has long been recognized. Bayesian "predictivism" centers on forecasting future observable variables (typically, a response of some system) in terms of the resulting predicting posterior distributions, and avoids traditional inferences on non-observable model parameters [54]. The focus of this paper is bayesian predictive techniques that deal with experimental data, obtained with the goal to optimize or improve a process or a product.

A specific level of system complexity that can be analyzed with a bayesian predictive approach is when the process response is multivariate, including the high dimensional case, common in current industrial data-rich environments. Related to the multivariate response process case, but significantly different from it, is the case of a process whose performance is not given by a collection of correlated scalar responses but is given instead by the shape of a continuous curve, i.e., a "profile" response system, in which an ideal profile is assumed to exist describing the best operation of the process [52].

The recognition of the usefulness of the bayesian paradigm has recently been pointed out by Tabora et al. [59] in the area of pharmaceutical process development. These authors recommend the bayesian predictive approach developed by Peterson et al. [41, 12, 42, 35] to consider the inherent uncertainties in process development and to quantify the risk in "Quality by Design" problems, as defined by the FDA. Even though the diffusion of bayesian methods to systems engineering, process analytics and chemometrics is still limited, a number of important problems have been studied. For instance, Nounou et al. developed bayesian latent variable model estimation methodologies [40, 39]. Chen et al. [8] propose to study traditional Chemometrics methods from a bayesian point of view. Mechanistic modeling has benefitted from bayesian formulations for the estimation of kinetic parameters from chemical reactions [45, 27]. Applications to data rectification [2], state estimation [9] and sensor fusion [20, 53, 3, 17, 57] have also been developed for dynamical systems. More recently, several bayesian process monitoring methodologies have also been developed, namely for multimode, non-linear and non-stationary processes [66, 29, 22, 31]. Particularly, the bayesian predictive approach is the basis of many contemporary machine learning methods, but explicit models and methods for the optimization of chemical processes are lacking.

In the optimization of a chemical process from data-based models obtained from experimental tests, it is of prime importance to compute the probability that a future experiment will reproduce the current result considering the different uncertainties involved [50]. There are uncertainties not only due to measurement errors but also in the model itself, or resulting from the type of experimental design used, that need to be considered during the optimization stage. Additional sources of variability in the process need to be consider as well. In the discrete-parts industry -but no so much in the process industry- a popular type of process optimization approach was initiated in the late 80s by the Japanese engineer G. Taguchi [60] who introduced the key concept of a *noise factor*. These are process or product variables that can be manipulated in a carefully controlled experiment, but once the

process operates regularly (or the product is in the marketplace) they cannot be controlled by the manufacturer anymore and, on the contrary, vary randomly. This situation leads to the *Robust Parameter Design* (RPD) problem (Taguchi referred to process variables as "parameters"), in which the goal is to find the settings of the controllable process/product variables that optimize the responses despite the uncontrolled variability introduced by the noise factors (see Figure 1), i.e., find an optimal solution that is robust with respect to noise factor variability. Both controllable and noise factors can be either continuous variables or categorical variables (taking values over a discrete set of levels). For an overview of Robust Parameter Design methods, see [1, 65, 37, 12]. As it will be shown below, all these sources of uncertainty can be handled with a bayesian predictive optimization approach.

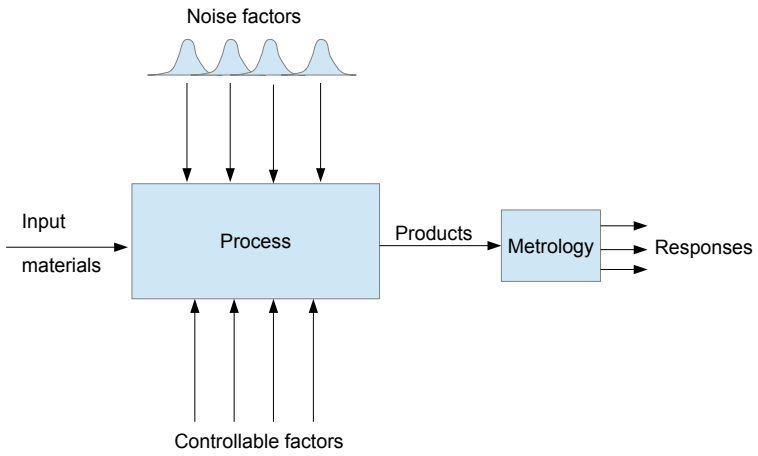

Figure 1: A schematic of a process as seen from the point of view of Taguchi's Robust Parameter Design (RPD). The goal in RPD is to find settings for the controllable factors (operating or design variables) that keep the responses of the process on the desired targets or goals in the presence of uncontrolled variability in the so-called noise factors. RPD requires process models, often obtained via experimental tests and statistical modeling.

In this paper, we present a review and some extensions of the bayesian predictive approach for the solution of optimization problems in the broader context of the process industry, including the RPD case. We focus on two types of empirical models: multiple response processes (perhaps with a very high dimensional response) and processes where the performance is described by some continuous curve or profile. These models have not been sufficiently covered in the literature, but are becoming increasingly relevant as more and more applications involve these type of high-dimensional responses (rather than the more common case of a high-dimensional predictor space as in machine learning applications). We illustrate the methods with 3 real-life applications in the process industry.

The remainder of the paper is organized as follows. We first contrast optimization approaches based on models fitted with classical (frequentist) statistical methods with a bayesian predictive approach. Next, two specific types of models, multivariate regression for multiple response processes and hierarchical mixed effects models for processes with a profile or curve response are presented, and their corresponding bayesian modeling and optimization is discussed, including the case of robust parameter design optimization. Here

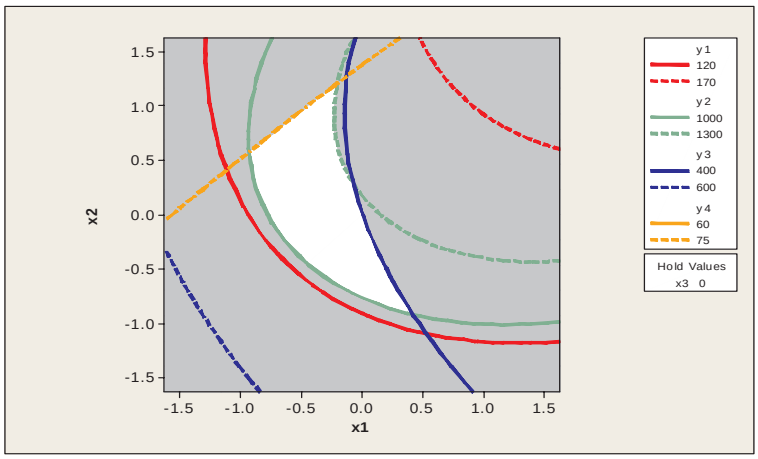

Figure 2: Overlaid contour plot of the four fitted responses in a tire tread experiment in the $x_1 - -x_2$ plane (keeping x_3 constant at 0.0). Unshaded region appears to be a "sweet spot" where to run the process as it seems to satisfy the response constraints. However, these are models for the mean responses, and neglect model uncertainty. See text.

a hierarchical mixed effects model previously used by del Castillo et al. [14] is adapted to the optimization of high dimensional responses via principal component analysis. Next, 3 examples are presented that illustrate the methods applied to industrial processes. Finally, we compare the hierarchical mixed effects model for a profile response to the increasingly popular probabilistic latent variable models used in Machine Learning, highlighting their different structure. MATLAB code is made available as supplementary material that accompanies this paper, and it implements all methods and examples presented here.

1.1 Limitations of process optimization based on classical statistical models

A common approach for process optimization based on experimental data is to fit regression models to responses and overlay contour plots to find a "sweet spot" where to run the process, a task facilitated by various popular statistical software packages. For a classical example, consider the experiment reported in [15] for the optimization of a car tire tread compound. The controllable factors were x_1, hydrated silica level, x_2, silane coupling agent level, and x_3, sulfur level. The four responses to be optimized and their desired ranges were: PICO Abrasion index, y_1, $120 < y_1$; 200% modulus, y_2, $1000 < y_2$; elongation at break, y_3, $400 < y_3 < 600$, and hardness, y_4, $60 < y_4 < 75$. Quadratic polynomial models were fitted to each of these responses from experimental data. Figure 2 shows overlaid contour plots (obtained with a popular statistical software package) which seem to imply it is safe to run the process within the "sweet spot" in the unshaded area.

However, it can be risky to follow this common practice. Suppose we fit a regression model to a process response of the form $\hat{Y} = g(x_1, x_2, .., x_k)$ from experimental data. Under the usual regression assumptions, the model fit is a prediction of the mean of Y, i.e., $\hat{Y} = \hat{E}[Y]$. Assume, for instance, that we wish to minimize the response by finding

operating conditions $\mathbf{x}^* = (x_1^*, ..., x_k^*)$ such that $Y(\mathbf{x}^*) = E[Y(\mathbf{x}^*)] \leq u$ where u is an acceptable upper bound for the response. It is important to note that all that $E[Y(\mathbf{x}^*)] \leq u$ guarantees for future times we observe the process at conditions $\mathbf{x}^*$ is that

$$P(Y(\mathbf{x}^*) \leq u) \geq 0.5$$

under the assumption of a symmetric distribution around the mean such as the normal. Likewise, in the case of two responses, if we find operating conditions $\mathbf{x}^*$ such that $E[Y_1(\mathbf{x}^*)] \leq u_1$ and $E[Y_2(\mathbf{x}^*)] \leq u_2$, all that is guaranteed at these conditions is that

$$P(Y_1 \leq u_1, Y_2 \leq u_2) > 0.25.$$

In general, suppose we have q responses in a process, and we fit corresponding regression models to $E[Y_1(\mathbf{x})], ..., E[Y_q(\mathbf{x})]$. If we then find either numerically or simply by overlapping contour plots of the responses the process conditions $\mathbf{x}^*$ such that $E[Y_1(\mathbf{x}^*)] \leq u_1$ and $E[Y_2(\mathbf{x}^*)] \leq u_2, ..., E[Y_q(\mathbf{x}^*)] < u_q$, all that is guaranteed at $\mathbf{x}^*$ is that

$$P(Y_1(\mathbf{x}^*) \leq u_1, Y_1(\mathbf{x}^*) \leq u_2, ..., Y_q(\mathbf{x}^*) \leq u_q) \geq 0.5^q$$

a probability that can indeed be very low. In practice, this probability bound is an overestimate, given that fitted models $\hat{Y}_j(\mathbf{x}) = \hat{E}[Y_j(\mathbf{x})]$ are used instead. Also, if the responses are positively correlated, then the lower bound will be higher and it may be easier to jointly optimize the system. If the responses are negatively correlated, the opposite would happen [42]. Peterson and Lief [43] document how in six real data industrial process optimization studies, including those in three published papers from the literature, the posterior probability of meeting the desired specifications actually varied from as low as 0.11, highlighting the danger of optimizing fitted functions for the mean response that neglect the uncertainties involved. In summary, optimizing regression models fitted to a set of responses and looking at overlaying contour plots:

- neglects model parameter uncertainty;
- can not provide a probability of assurance or "reliability" about whether future responses at given process settings $\mathbf{x}$ will satisfy process specifications;
- neglects the covariance between the responses during the optimization step, even if models were fitted via classical multivariate regression, which does consider these covariances during the estimation step.

1.2 Advantages of the bayesian predictive approach for process optimization

Billheimer [4] has recently advocated the bayesian predictive approach for statistical inference. A natural way to optimize any process from a quality and reliability standpoint is to maximize the probability of conformance of the predicted responses to their specification limits. Bayesian predictive models and their use in industrial process optimization

were pioneered by Peterson [41] in pharmaceutical applications (see also [12] for a detailed presentation). As Billheimer [4] indicates, "A scientist is interested in the *probability* that a future experiment will reproduce the current result". Likewise, in process optimization, an engineer is interested in the probability that the future operation of the process will result in the optimal performance that settings deduced from past experiments seems to indicate.

In the pharmaceutical sector, the ICH Q8 and Q11 guidelines [18] for industry promoted the concept of a "design space", defined as the combination of input variables and process parameters that have been demonstrated to provide assurance to quality". Using the methods presented here it is possible to find a design space for a process that with *known probability* is predicted to achieve particular goals if run inside the space (see [58], and for an instance, see example 2 below).

In summary, the advantages of the bayesian predictive optimization approach compared to optimizing models fitted classically are:

- it provides a probability estimate that future responses at given operating conditions $\mathbf{x}$ will satisfy the desired process specifications;
- it considers the uncertainty in the model parameters, not only in the measurements;
- it takes into account the correlation between the responses during the optimization (not only during model building);
- it permits to include additional sources of variability such as "noise factors" in the Taguchi sense;
- it can be extended to different types of responses, as reviewed below.

For further discussion of disadvantages of the classical approach for response surface optimization and the advantages of a bayesian approach, see [12], Chapter 12.

2 Bayesian inference for multiple response and profile response processes

2.1 Central role of the predictive density in bayesian process optimization

Regardless of the model that is most appropriate for a response Y, the object of interest for bayesian optimization is the *posterior predictive density* (or predictive density, for short) of the response, $f(\tilde{Y}|x, \text{data})$ where "data" denotes all the experimental data available, including controllable and noise factors and the corresponding observed response values, and the tilde on Y denotes a *future* response value not yet observed. The predictive density contains all the relevant information about a response that is needed to make inferences

about it. Given a statistical model with parameters θ, the predictive posterior density is defined as:

$$f(\tilde{Y}|x, \text{data}) = \int p(Y|\theta, x, \text{data}) p(\theta|\text{data}) d\theta \tag{1}$$

where $p(Y|\theta, x, \text{data})$ is the likelihood function and $p(\theta|\text{data})$ is the posterior distribution of the model parameters. Depending on the model for the response Y, the integral in equation (1) may have a closed form expression (this is the case of the multiple response regression case, see below). In case it does not have a closed form, Markov Chain Monte Carlo (MCMC) techniques can be used to find the posterior of the parameters and, with it, the posterior predictive density. Either informative or non-informative priors on θ can be used, but informative priors are usually difficult to justify and it is usually better to work with non-informative priors, see [23, 13, 11, 12].

For a specific illustration, let us consider the posterior predictive density for a basic linear regression model $Y = \beta_0 + \beta_1 x_1 +, ..., +\beta_k x_k + \varepsilon$ with $p = k + 1$ parameters. In this case, the posterior predictive density $p(\tilde{Y}|Y, \mathbf{x})$ is given by a Student t distribution with $N - p$ degrees of freedom, mean $\mathbf{x}'\hat{\boldsymbol{\beta}}$ (where $\hat{\boldsymbol{\beta}}$ is the ordinary least squares estimate of $\boldsymbol{\beta} = (\beta_0, ..., \beta_k)' = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}$) and variance $S^2(1 + \mathbf{x}'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x})$. Here, S^2 is the sample variance and $\mathbf{X}$ is an $N \times p$ matrix with one column per model term and one row per experimental test. From these results, it is easy to verify that the *variance* of the predictive density is a function of:

- the amount of data (N, as seen in matrix $\mathbf{X}$ and the sample variance S^2);
- the inherent noise due to measurement error (estimated via S^2);
- how well the model fits (as defined by the columns of $\mathbf{X}$);
- the model parameter uncertainty, determined in good part by the experimental design used, present in matrix $\mathbf{X}$;
- the variability of the noise factors $\mathbf{x}_n$ where $\mathbf{x} = (\mathbf{x}_c, \mathbf{x}_n)$ are all factors in the experiment (controllable and noise factors, respectively);
- our ability to control the noise factor variability thanks to the presence of interaction terms in the model between $\mathbf{x}_c$ and $\mathbf{x}_n$ variables. If such control×noise interactions do not exist in the model, there is no way to solve the RPD problem [12].

2.2 Bayesian optimization of multi-response linear regression models

Let $Y_1, Y_2, ..., Y_J$ be J responses of interest in a process. To investigate the input-output relations of the process, an experiment is designed and conducted by varying k controllable factors $x_1, ..., x_k$ over N different test conditions or "runs". Assume we fit a linear regression model to each of the J responses (same model for all responses assumed in this section) each containing q parameters. If the experiment is "large" enough in the sense that:

$$J < N - q + 1 \tag{2}$$

then a standard *multivariate* linear regression model can be fitted, written in matrix notation as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Gamma} + \mathbf{E} \tag{3}$$

where $\mathbf{Y}$ is a $N \times J$ matrix containing the observations of the J responses in each of the N experimental runs (combinations of controllable and noise factors). Also, $\mathbf{X}$ is a $N \times q$ design matrix, with all the controllable and noise factors expanded in model form according to the same q-parameter model. $\mathbf{\Gamma}$ is a $q \times J$ matrix containing all the q parameters for each response $j = 1, 2, ..., J$ (note the same model form is assumed for all responses), and $\mathbf{E} \equiv [\varepsilon_{ij}]$ is a $N \times J$ matrix of random errors which are assumed to be probabilistically described, as follows:

$$\begin{array}{rcll} \boldsymbol{\varepsilon}_{i.} & \sim & \mathbf{N}_J(\mathbf{0}, \mathbf{\Sigma}) & \forall i = 1, \ldots, N \\ \boldsymbol{\varepsilon}_{.j} & \sim & \mathbf{N}_N(\mathbf{0}, \sigma_j^2 \mathbf{I}_N) & \forall j = 1, \ldots, J. \end{array}$$

That is, the model assumes that errors between responses from the same experiment (a row of $\mathbf{E}$) can be correlated, but it also assumes that the errors along a column of $\mathbf{E}$ (errors for the same response Y_j for different experimental runs i) are independent random variables. Under this model, a $J \times 1$ vector $\mathbf{y}$, containing a single, not yet observed, vector of responses, for given levels of the controllable and noise factors $\mathbf{x}$, is assumed to follow the model:

$$\mathbf{y} = \mathbf{\Gamma}' \mathbf{f}(\mathbf{x}) + \boldsymbol{\varepsilon} \tag{4}$$

where $\mathbf{f}(\mathbf{x})$ is a $q \times 1$ vector containing the values of the controllable and noise factors $\mathbf{x} = (\mathbf{x}_c, \mathbf{x}_n)$ at which the prediction is desired (expanded in model form, same form as used in the columns of $\mathbf{X}$ and $\mathbf{\Gamma}$) and $\boldsymbol{\varepsilon}$ has the same distribution as a row of $\mathbf{E}$, i.e., a $N(\mathbf{0}, \mathbf{\Sigma})$ distribution.

Fortunately, the posterior predictive density of model (4) is available in closed form. Here one can utilize non-informative priors, to avoid heavily weighted priors that are hard to justify. If condition (2) does not hold because the response is very high dimensional, the hierarchical mixed effects model described further below, originally proposed for the optimization of profile responses, can be used as well as we explain in example 1.

Under the classical non-informative joint prior for $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ in equation (3), it is well-known (see, e.g., see [46], pp. 136 or [12]) that the bayesian predictive density for a new response vector $\mathbf{y}$ is given by a J-dimensional t distribution with $\nu = N - q - J + 1$ degrees of freedom:

$$f(\tilde{\mathbf{y}}|\mathbf{x}, data) = \frac{\Gamma\left(\frac{\nu+J}{2}\right)}{(\pi\nu)^{J/2}\Gamma\left(\frac{\nu}{2}\right)} \sqrt{|\mathbf{H}|} \left\{ 1 + \frac{1}{\nu} \left(\tilde{\mathbf{y}} - \widehat{\mathbf{B}}'\mathbf{x} \right)' \mathbf{H} \left(\tilde{\mathbf{y}} - \widehat{\mathbf{B}}'\mathbf{x} \right) \right\}^{-\frac{\nu+J}{2}} \tag{5}$$

This is denoted $\mathbf{T}_J^{\nu}(\mathbf{a}, \mathbf{b})$, where $\mathbf{a}$ is the mean vector and $\mathbf{b}$ is the variance matrix. That is, $\tilde{\mathbf{y}}|\mathbf{x}, data \sim \mathbf{T}_J^{\nu}\left(\widehat{\mathbf{\Gamma}}'\mathbf{x}, \frac{\nu}{\nu-2}\mathbf{H}^{-1}\right)$, where $\widehat{\mathbf{\Gamma}}$ is the ordinary least squares (OLS) estimator of $\mathbf{\Gamma}$, and $\mathbf{H}$ is given by:

$$\mathbf{H} = \left(\frac{\nu}{N-q} \right) \frac{\widehat{\mathbf{\Sigma}}^{-1}}{1 + \mathbf{x}' \left(\mathbf{X}'\mathbf{X} \right)^{-1} \mathbf{x}}$$

where $\widehat{\mathbf{\Sigma}}$ is the usual Maximum Likelihood Estimator (MLE) of $\mathbf{\Sigma}$.

Based on the bayesian multivariate regression formulation above, Peterson [41] proposed a method to conduct multiple-response optimization that accounts for the uncertainty in

the parameters of the model and for any correlation present between the responses. His method assumes no noise factors, hence $\mathbf{x} = \mathbf{x}_c$ and consists in solving

$$\begin{aligned} max \quad & p\left(\mathbf{x_c}\right) = P(\tilde{\mathbf{y}} \in A | \mathbf{x_c}, data) \\ subject\ to: \quad & \\ & \mathbf{x_c} \in R_c \end{aligned}$$

where A is a desired specification region for the J responses and R_c is a feasible region for the controllable factors. In its simplest form, both A and R_c are given by upper and lower bounds.

To obtain $p\left(\mathbf{x_c}\right)$ through equation (6), the predictive density $f(\tilde{\mathbf{y}}|\mathbf{x}_c, data)$ needs to be integrated (numerically) over the region A:

$$p\left(\mathbf{x_c}\right) = \int_A f(\tilde{\mathbf{y}}|\mathbf{x_c}, data)\, d\tilde{\mathbf{y}}. \tag{6}$$

For a multivariate regression model, this can easily be done by Monte Carlo simulation of a multivariate t distribution, as explained in Appendix A.

If noise factors are present in the system and affect the responses (RPD case), a bayesian approach that provides solutions which are robust both to the noise factor variability and to the uncertainty in the model parameters was proposed by Miro et al. [35] as an extension to Peterson's method. It consists in solving the following optimization problem,

$$\begin{aligned} max \quad & p(\mathbf{x_c})_{\text{RPD}} = \int P(\tilde{\mathbf{y}} \in A | \mathbf{x}, data) f(\mathbf{x}_n) d\mathbf{x}_n \\ subject\ to: \quad & \\ & \mathbf{x_c} \in R_c. \end{aligned} \tag{7}$$

That is, after obtaining $p(\mathbf{x})$ for fixed $\mathbf{x} = (\mathbf{x_c}, \mathbf{x_n})$, a second integration is performed over the (assumed known) distribution of the noise factors, $f(\mathbf{x}_n)$, to obtain the probability of conformance to specifications (also called the "reliability" of the process [41]) for the RPD problem, $p(\mathbf{x_c})_{\text{RPD}}$. Miro et al. [35] considered only normally distributed noise factors. We extend this below to the case of Bernoulli(p) and generalized Bernoulli($p_1, p_2, ..., p_L$) noise factors, given that in many cases *categorical* noise factors, with L levels each, are uncontrolled and can not be treated as continuous random variables. The MATLAB code provided (see supplementary materials) implements all integrations and the optimization needed to solve problem (7).

If the optimal probability after solving (7) is high, then there is a high degree of assurance that the specification region A will be achieved, despite the noise factor and other sources of variability. We would then have a robust solution to the RPD problem. The solution will also consider the uncertainty in the model parameters and measurement error, since it is based on the predictive density of the response. Model parameter uncertainty is linked to the specific type of experimental design used, which determines matrix $\mathbf{X}$.

2.3 Bayesian predictive optimization based on other response models

The multivariate regression model presented earlier assumes all J responses follow the same exact model form with p parameters each. This may constitute a limitation in practical

applications. Therefore, this model and its use in process optimization has been extended in [44] to the case where each response can adopt a different linear model, the so-called "SUR" (seemingly unrelated regression) case.

The optimization depends on the models used. There are situations where more than one model fits the data reasonable well, but their corresponding optima lie in different regions of the controllable factor space. Rajagopal and del Castillo [47] expand the bayesian optimization problem to the case a set of models $\{M_i\}$ fit the response adequately, solving instead the problem:

$$\max_{\mathbf{x}_c \in R_c} \quad \sum_{\text{all}i} p(\tilde{Y}|M_i, \text{data}, \mathbf{x}) d\tilde{Y} p(M_i|\text{data})$$

where $p(M_i|\text{data})$ is the posterior density of each model i. The model-averaged solution found is then robust with respect to variation in the true model describing the response. This methodology was extended by Ng [38] to the multiple response case, allowing a user to define alternative models for each response.

The bayesian models presented above assume normally distributed data. To robustify this assumption, Rajagopal et al. [48] considered instead the noise in the model as originating from a Student t distribution. The models above also assume a "steady state" input-output behavior of the process. It is possible to extend bayesian optimization ideas to the dynamic case where some of the variables are lagged. Vanli and del Castillo [62] consider bayesian RPD optimization of a process based on a linear regression model in which the noise factors randomly vary according to a time series model, and frequent reoptimization is necessary to adapt to the varying noise factor state.

2.4 Bayesian optimization of response profiles

In cases where the performance of a process is given by a curve or "profile", as opposed to a set of scalar responses, a different model is necessary, and here we review a useful hierarchical mixed effects model for profile responses originally presented in [14]. Let us assume the response $Y(s)$ can be observed at several fixed values of an auxiliary variable s which we will refer to as the "locations" $s_1, s_2, ..., s_J$. For each experimental run i ($i = 1, ..., N$) where controllable and noise factors $\mathbf{x}_i = (\mathbf{x}_c, \mathbf{x}_n)_i$ have been tried in a designed experiment, a complete profile or function is observed consisting of J points along the curve $Y(s)$. Thus, rather than observing a continuous curve $Y_i(s|\mathbf{x}_i)$, we observe the discrete response

$$Y_{ij} = g(\mathbf{x}_i; s_j) + \varepsilon_i(s_j), \; i = 1, ..., N, j = 1, ..., J, \tag{8}$$

where g is some function to be specified/estimated and $\varepsilon_i(s_j)$ is a random error, which can depend on the location s_j.

The multivariate regression approach could be used again treating the different response values at different locations s_j as different responses. If the number of experiments is large such that condition (2) holds, this would be feasible. However, while this approach provides predictions at the predefined locations values s_j, it is not possible to *interpolate* with this

model, since the locations s_j are not used explicitly in the model. That is, it is not possible to make predictions at profile locations s_j other than those observed during the experiment.

When the number of points per profile J is large ($J \geq N-q+1$), the previous approach cannot be applied. An alternative then is to use informative priors and optimize the posterior predictive density, but very informative priors are difficult to justify in general. A more general approach is needed. Del Castillo et al. [14] propose to model a profile response using a hierarchical Bayes approach. In a first stage, the curves or profiles are modeled as a regression function of the locations s:

$$Y_{ij} = g(\mathbf{x}_i; \theta_i^{(1)}, \theta_i^{(2)}, ..., \theta_i^{(q)}; s) + \varepsilon_i.$$

The *parameters* $\{\theta_i^{(k)}\}$ are then modeled as a function of controllable and noise factors $(\mathbf{x}_c, \mathbf{x}_n)$ at a second stage. Thus, modifying the controllable factors affects the first stage parameters, which in turn modify the form of the curve response.

The model in [14] is

$$\mathbf{y}_i = \mathbf{S}\boldsymbol{\theta}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim N_J(\mathbf{0}, \boldsymbol{\Sigma}), \tag{9}$$

and second stage model given by

$$\boldsymbol{\theta}_i = \mathbf{B}\mathbf{f}(\mathbf{x}_i) + \mathbf{w}_i, \quad \mathbf{w}_i \sim N_p(\mathbf{0}, \boldsymbol{\Sigma}_w) \tag{10}$$

for $i = 1, .., N$, where $\mathbf{y}_i$ is a $J \times 1$ vector containing the observations along profile i, $\mathbf{S}$ is a $J \times p$ matrix of regressors for fitting the stage 1 model (here the regressors will be functions of the locations s), $\boldsymbol{\theta}_i$ is a $p \times 1$ vector of stage 1 parameters, and $\mathbf{B}$ is a $p \times q$ matrix of parameters –the stage 2 parameters– containing the effects of $\mathbf{x}_i = (\mathbf{x}_c, \mathbf{x}_n)_i'$, the experimental conditions in run i. The notation $\mathbf{f}(\mathbf{x}_i)$ indicates as before that $\mathbf{x}_i$ is expanded in model form to include all q terms in stage 2. Thus, each element of the parameter vector $\boldsymbol{\theta}_i$ is assumed to be modeled adequately by a model containing q parameters.

The hierarchical model (9-10) can be written as the single linear model:

$$\mathbf{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \mathbf{S}\mathbf{w}_i + \boldsymbol{\varepsilon}_i \tag{11}$$

where $\boldsymbol{X}_i = \mathbf{f}(\mathbf{x}_i)' \otimes \mathbf{S}$ is a $J \times qp$ matrix and $\boldsymbol{\beta} = vec(\mathbf{B})$ is a $qp \times 1$ vector. This is a model widely used in Biostatistics, in particular, in longitudinal growth curve analysis (see [32, 19]) where it is called a linear *mixed effects* model, since the term $\mathbf{S}\mathbf{w}_i$ is stochastic but the term $\boldsymbol{X}_i\boldsymbol{\beta}$ is not (i.e., it has a "fixed" effect). Notice how if using this model, we would be fitting pq parameters, compared to Jq parameters that would be needed if the multi-response approach would be used considering each of the J locations different responses. Thus, if J is large compared to p, this would represent a more parsimonious model. This is also an alternative for the case $J > N - q + 1$, when the single stage multi-response approach of the previous section cannot be applied. However, model (11) requires the estimation of a $J \times J$ covariance matrix, unless simplifying assumptions are made. Hierarchical approaches based on (9-10) explicitly utilize the information about the locations (contained in matrix $\mathbf{S}$). This is not the case in the single stage multiple response approach presented earlier.

As discussed by [63], model (11) is unnecessarily restricted, since the design matrices for the fixed effects term and for the random effects term are linked. For this reason, del Castillo et al. [14] suggested to use *different* matrices $\mathbf{S}$ and $\mathbf{S}^*$ in (11), keeping the original $\mathbf{S}$ matrix with the location information for the fixed effects term but selecting $\mathbf{S}^*$ to match the observed within-profile covariance. The model is therefore:

$$\mathbf{y}_i = \boldsymbol{X}_i\boldsymbol{\beta} + \mathbf{S}^*\mathbf{w}_i + \boldsymbol{\varepsilon}_i \tag{12}$$

where we still have that $\boldsymbol{X}_i = \mathbf{x}_i^{(m)'} \otimes \mathbf{S}$. Their method to select $\mathbf{S}^*$, however, frequently results in $\mathbf{S}^* = \mathbf{S}$. A better approach, which we fully explain in example 3 below (section 3.1.1), is to select $\mathbf{S}^*$ using Probabilistic Principal Components (PPCA) analysis [61].

The parameters of model (12) are $\boldsymbol{\beta}, \{\mathbf{w}_i\}, \boldsymbol{\Sigma}_w$, and σ^2. Bayesian inference for this model has been studied by some authors ([33, 10, 14]) and requires Markov Chain Monte Carlo (MCMC) sampling, since the joint posterior of these parameters is not a known distribution in closed form. Appendix B gives the full conditional distributions of each parameter needed in the Gibbs sampling algorithm that yields the joint posterior of the model parameters.

The probability of conformance to a set of specifications $p(\mathbf{x_c})_{\text{RPD}}$ is maximized once the predictive density of the response along the profile, $\mathbf{y}|\mathbf{x}, data$ is obtained via the MCMC scheme in Appendix B. As described in that appendix, the Markov chain is run once until convergence and sampled within the optimization whenever a value of $\mathbf{y}|\mathbf{x}, data$ is needed. The MCMC sampling scheme and the optimization needed are implemented in the MATLAB code we are providing (see supplementary materials).

3 Bayesian predictive optimization in the process industry: some examples

In the examples below we fit either the bayesian multivariate regression model or the bayesian hierarchical mixed effects model, depending on the type of response to be handled. Each model has a variety of model diagnostics that should be consulted. For diagnostics pertaining to the hierarchical mixed effects model, see ref. [14], where several residual plots and normality tests are presented. These were all checked in the examples that follow, and are omitted for brevity.

3.1 Example 1: Quantification of analytes impacting wine aroma

The following example illustrates the RPD optimization method using multivariate regression (the case when $J > N - q + 1$). The optimization of analytical instrumentation is a major activity in research and industrial laboratories. It is fundamental to take the most out of the high capital and operational costs involved including operator time. Quite often the measurement system should address multiple targets, becoming a multi-response process. For instance, Reis et al. [49] studied the optimization of a headspace solid-phase

microextraction (HS-SPME) method, which is a state of the art extraction methodology for the analysis of volatile chemical compounds in liquid samples. The goal is to optimize the quantification of analytes impacting wine aroma. They reported the results from a 7 factor experimental design study consisting of $N = 18$ runs, where the responses were the chromatographic responses of 9 compounds establishing wine flavor (peak area of each compound in the chromatogram). These 9 compounds are: isobutyric acid, butyric acid, isovaleric acid, valeric acid, hexanoic acid, octanoic acid, nonanoic acid, decanoic acid and dodecanoic acid.

In order to apply bayesian optimization, we initially treat the 9 chromatographic responses as a 9-dimensional response vector. The goal of the HS-SPME extraction method is to detect the compounds, so the larger the peak areas the better. The responses were standardized and the controllable factors coded into the (-1,1) convention. Out of the seven factors reported by [49] (fiber coating, pre-incubation time, extraction time, extraction temperature, headspace sample volume, agitation during extraction and ethanol content) two of them, pre-incubation time and agitation seem not to have any effect on the responses and were eliminated from the study. The final set of the factors were x_1=VOL (headspace sample volume, ranging from 5 to 10 ml.), x_2=ETI (extraction time, ranging from 15 to 20 min.), x_3=ETE (extraction temperature, ranging from 40 to 55 °C.), x_4=EC (Ethanol content, ranging from 4.5 to 18%)and x_5=F (type of fiber, a categorical factor with labels {L1-PA,L2-DVB}.

In the following analysis, we treat x_5 as a noise factor with a Bernoulli(0.5) distribution, implying that we wish to find settings in the (x_1, x_2, x_3, x_4) controllable factors that make the chromatographic areas large with high probability *regardless* of the type of fiber used.

Since the goal is to maximize the responses and they are standardized, we set the bounds $L_i = 0.5$ and $U_i = 3$ for all responses $i = 1, ..., 9$. We define the tolerance or specification region for the responses A as the hyperbox defined by the intervals (L_i, U_i). We then wish to maximize the joint posterior predictive probability that the 9 chromatographic responses are all inside their bounds. The model fit to each response was:

$$Y = \beta x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_{12} x_1 x_2 + \beta_{15} x_1 x_5 + \beta_{34} x_3 x_4 + \beta_{44} x_4^2$$

(note there is no intercept as the Y's were standardized). The bayesian multivariate regression model (3) was fit and the RPD optimization problem (7) was solved. Tables 1-2 show the maximum probabilities of satisfying the given specifications and the resulting solution $\mathbf{x}_c^*$, respectively. Figure 3 shows the point prediction of the optimal responses (mean vector of the posterior predictive density) and one-sigma standard errors. The errors are evidently too large compared to the specifications.

$p(\mathbf{Y}(\mathbf{x}_c) \in A\|\text{data})$	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
0.110	0.279	0.557	0.496	0.537	0.524	0.663	0.678	0.936	1.000

Table 1: Estimated predictive posterior probabilities of satisfying the joint bounds A and each interval $A_i = (L_i, U_i)$. Here, $p_i = p(L_i \leq Y_i \leq U_i|\text{data})$ using a multivariate regression analysis.

	VOL=x_1	ET1=x_2	ETE=x_3	EC=x_4
x^* (coded)	0.950	0.993	-0.081	0.827
x^* (uncoded)	9.87	39.61	47.63	16.83

Table 2: Optimal settings for Wine experiment original experiment ($N = 18$, $\nu = 1$), multivariate regression analysis.

The optimal solution (uncoded), $\mathbf{x}_c = (9.87, 39.61, 47.63, 16.83)$ coincides closely to what was found by [49]. Note that there is no run similar to this solution in the original experimental design, so simply "picking the winner" from the list of experimental trials would not have resulted in an optimal solution.

The overall probability of satisfying the 9 bounds at $\mathbf{x}^*$ is quite low, only 0.110. This indicates either problems with the amount of data available or the possibility that the bounds (L, U) considered in the analysis were too strict. We conducted a *preposterior analysis* [41] to discern if more data would have resulted in better probabilities $p(\mathbf{Y}(\mathbf{x}_c) \in A|\text{data})$ or if the low probability is due to unrealistic bounds. We repeated the optimization by duplicating the design and the response data ($N = 36$). The results are shown in Figure 4, and as it can be seen, the standard errors are much smaller. The optimal solution x_c^* changed little, and the overall probability of conformance went up but only to 0.326. Furthermore, artificially increasing the number of runs does not increase this probability any further, an indication that either the bounds are unrealistic or the response information collected from the experimental design is limited.

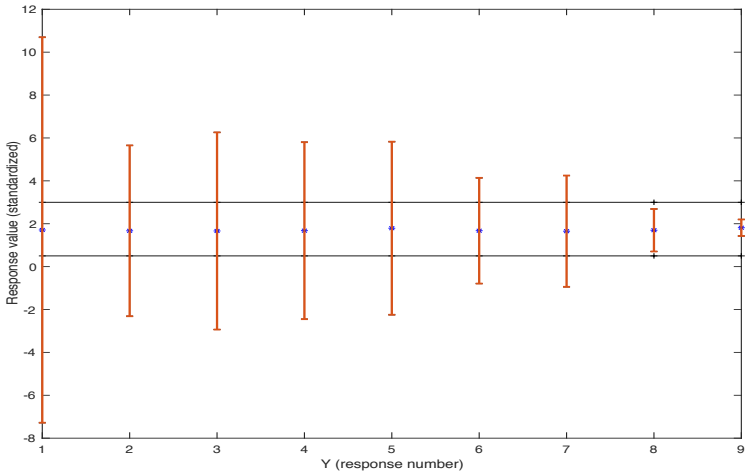

Figure 3: Predicted responses (blue stars) at the optimal settings with one std. deviation error bars, using the original experimental design, wine aroma experiment, $N = 18$ and $\nu = 1$ degree of freedom. Optimization using the multivariate regression model.

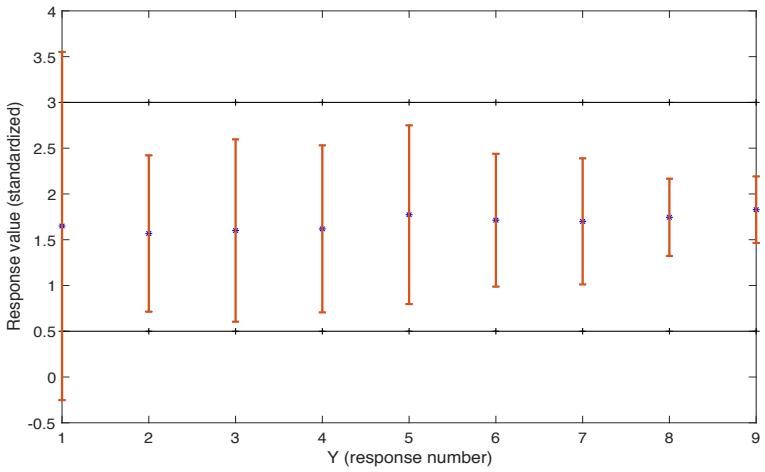

Figure 4: Preposterior analysis, predicted responses (blue stars) at the optimal settings in the wine aroma experiment, with one std. deviation error bars assuming the original design was duplicated, $N = 36$ and $\nu = 19$ degrees of freedom (note different y-scale compared to figure 3). Optimization based on multivariate regression model.

3.1.1 Reanalyzing the wine aroma experiment using the hierarchical mixed effects model as a response dimensionality reduction method

The large standard errors of the predictive density, and the low predictive probability of conformance to the specifications occur because in reality the 9 chromatographic responses are highly correlated so that their effective dimension is much lower than 9. Therefore, we may follow the modeling approach proposed in [49] and first conduct a principal component analysis (PCA) analysis on the 9 responses, which identifies the first two principal components (PC) as explaining more than 95% of the variability in the responses. We then define the matrix $\mathbf{S}$ in the hierarchical model (9-10) by these two principal components loadings:

$$\mathbf{S} = \begin{pmatrix} 0.3304 & 0.3797 \\ 0.2670 & 0.6036 \\ 0.3486 & 0.2596 \\ 0.3534 & -0.0873 \\ 0.3361 & -0.3424 \\ 0.3331 & -0.3563 \\ 0.3393 & -0.3385 \\ 0.3219 & -0.1733 \end{pmatrix}$$

Note that the first PC is proportional to the average of the nine responses while the second PC is similar to a contrast between analytes with different chain lengths. We now solve the RPD optimization problem with the same bounds (for the 9 responses) as before. Tables 3-4 show the optimal probabilities of conforming to the same specifications presented before and the optimal $\mathbf{x}_c$ settings. The overall predictive probability of satisfying the response specifications is much higher, 0.857 (see Figure 5 for a plot of the predicted optimal responses, which shows a much more concentrated optimal predictive density around its

mean). The original problem had responses highly correlated, effectively diffusing the probability over a space that was too high dimensional with respect to the subspace where the data really lied. Using a smaller number of PCs (2) captures better the underlying probability distribution. Note that the optimal solution $\mathbf{x}_c^*$ did not change much from model to model. What was gained was a higher degree of *assurance* that the optimal process will meet its specifications.

$p(\mathbf{Y}(\mathbf{x}_c) \in A \vert \text{data})$	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
0.857	0.998	0.991	0.997	0.992	0.970	0.935	0.935	0.951	0.986

Table 3: Estimated predictive posterior probabilities of satisfying the joint bounds A and each interval $A_i = (L_i, U_i)$. Here, $p_i = p(L_i \leq Y_i \leq U_i \vert \text{data})$ using the mixed effects model based on a preliminary PCA of the response data.

	VOL=x_1	ETI=x_2	ETE=x_3	EC=x_4
x^* (coded)	0.571	0.995	0.091	0.9984
x^* (uncoded)	8.92	39.93	48.18	17.98

Table 4: Optimal settings for Wine experiment original experiment ($N = 18$), mixed effects model based on a preliminary PCA of the response data.

In this analysis based on the mixed effects model, as well as in the next two examples below, the design matrix $\mathbf{S}^*$ used in model (12) was estimated using Probabilistic PCA (Appendix C) applied to the residuals $\mathbf{r}_i$ of the model

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{r}_i$$

where $\mathbf{r}_i = \hat{\mathbf{Y}}_i - \mathbf{Y}_i$, such that $\mathbf{r}_i = \mathbf{W}\mathbf{z} + \boldsymbol{\varepsilon}$ where we make $\mathbf{W}$ in equation (13) equal to $\mathbf{S}^*$ in (12), see Appendix C.

3.2 Example 2: Optimization of the tensile stiffness orientation profile in a paper machine

Reis and Saraiva [52] considered the prediction of a profile response in a paper manufacturing facility. The response of interest is the tensile stiffness orientation (TSO) angle profile across a paper sheet produced in a paper machine. The TSO orientation is closely related to the fiber orientation angle that has a major effect in paper mechanical and dimensional properties. For instance, higher angles tend to favor diagonal curl modes, which can be very detrimental to printing processes, causing frequent jams, loss production and reduced operational efficiency. The profile of TSO across the paper machine can be controlled by manipulating several process variables (experimental factors), such as x_1=VJ (jet velocity, ranging from 800 to 850 m/min), x_2=(A, slice opening, ranging from 40 to 80%), and x_3=(VR, manifold recirculation, ranging from 40 to 80%). The wire velocity (VW) of

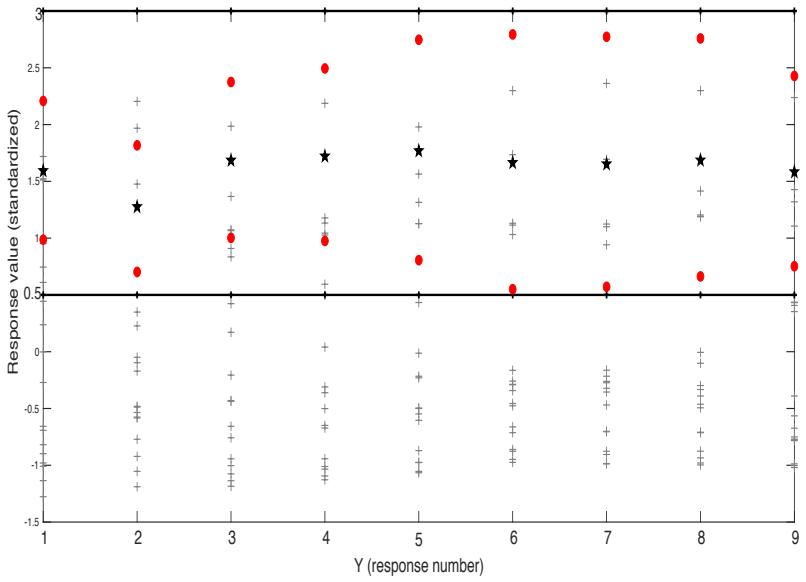

Figure 5: Optimal responses (stars) and 95% prediction interval for the responses in the wine aroma experiment under the solution obtained using the mixed effects model based on the first 2 principal components of the 9 chromatographic responses.

the machine was fixed at 750 m/min (this sets the production pace, which was assumed to be kept fixed). For operational reasons, |VW-VJ| $>$ 50 a condition that was always maintained during experimentation, as this is a machine requirement to balance paper properties (surface quality, dimensional and mechanical properties).

Here we model the observed TSO profiles with the hierarchical mixed effects model (9-10) with the first stage model given by:

$$Y_i = \theta_0 + \theta_1 s + \theta_2 \sin\left(\frac{2\pi}{J} s\right) + \varepsilon_i, \quad i = 1, ..., N$$

and second stage model equal to:

$$\begin{aligned} \theta_l &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \beta_{11} x_1^2 + \beta_{22} x_2^2 \\ &+ \beta_{33} x_3^2 + w_l, \quad l = 1, 2, ..., N \end{aligned}$$

A D-optimal, $N = 20$ run designed experiment was conducted, which gives adequate degrees of freedom to fit the quadratic polynomial model in stage 2. The object of the experimentation is to find conditions on the machine that make the TSO profiles as flat as possible. We therefore define bounds at $U_j = 3$ and $L_j = -3, j = 1, ..., J = 10$. The 20 observed TSO profiles were obtained at 10 positions over the machine manifold and are displayed, together with the predicted profiles (from the mean of the posterior predictive density) in Figure 6. The optimal TSO profile is displayed in Figure 7. The optimal solution $\mathbf{x}^* = \mathbf{x}_c$ (no noise factors were considered here) that maximizes the posterior predictive probability that the profile jointly meets its specifications is shown in Table 5

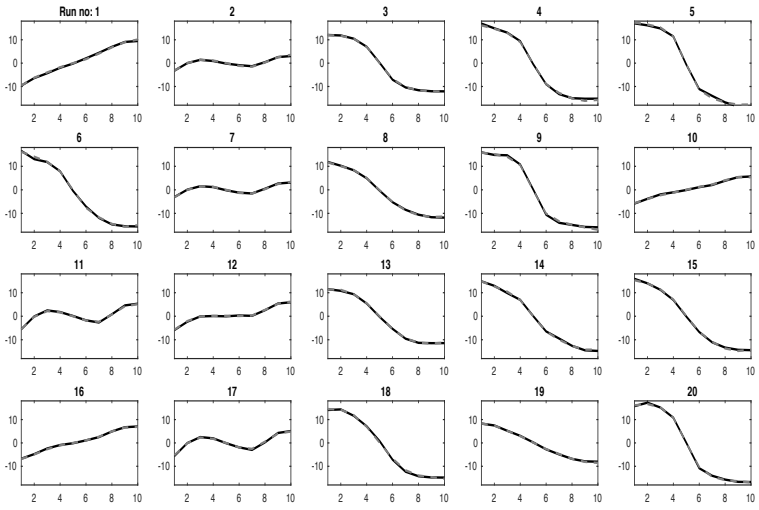

Figure 6: Dark lines: observed TSO profiles in each of the $N = 20$ experimental runs. Lighter, dashed lines: predicted TSO profiles (mean of the posterior distribution). The predictions are so close to the actual they are hard to see.

	VJ=x_1	A=x_2	VR=x_3
x^* (coded)	0.898	0.995	-0.732
x^* (uncoded)	847.4	79.90	45.35

Table 5: Optimal settings for Pulp and Paper experiment.

and the corresponding probabilities are shown in Table 6. Clearly, this is a very good solution, as TSO variation across the paper sheet is very low.

The profile main effects plot reveals how a low setting for the recirculation compensates with the average trend the profiles have, resulting in a flatter TSO response (see Figure 8).

We can in addition find a "design space" for this process by evaluating the probability of meeting the specifications $p(\mathbf{Y} \in A|\text{data}, \mathbf{x})$ at a variety of $\mathbf{x}$-points [43, 58] . This is feasible of course when the number of controllable factors is low. Figure 9 shows a contour plot of the probability function evaluated at different lip opening (x_3) and recirculation (x_4) values keeping the jet velocity fixed. A confirmation run was made at these settings. Figure 10 shows the actual profile.

3.3 Example 3: Optimization of a drug release stability profile.

Silva et al. [56] applied Quality-by-Design techniques to find the root cause for the observed slower drug release in orodispersible films during storage. The response is the drug release time profile, which should follow a given reference profile within certain bounds and should also be stable along time upon repetition of the same drug release trial. The following conditions were analyzed as experimental design factors to manipulate: x_1= RT = Room

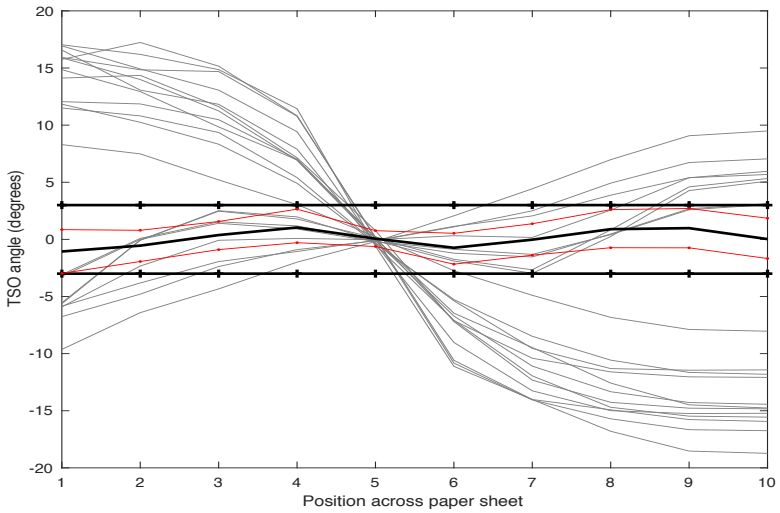

Figure 7: Light lines: observed TSO profiles, dark line: optimal predicted TSO profile, red lines: the 5 and 95% predicted percentiles of the optimal TSO posterior distribution, crossed dark lines: upper and lowed specifications for the TSO profile.

$p(\mathbf{Y}(\mathbf{x}_c) \in A \vert \text{data})$	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
0.898	0.983	0.997	0.939	1.000	0.977	0.994	0.939	0.935	0.962

Table 6: Estimated predictive posterior probabilities of a TSO profile satisfying all bounds $A_i = (L_i, U_i)$ simultaneously and at each point s. Here, $p_i = p(L_i \leq Y(s_i) \leq U_i|\text{data})$.

temperature, varying from 17 to 25 °C, x_2 = RH = room humidity, varying from 30 to 62 %, x_3 = DT = drying temperature, varying from 40 to 60 °C, x_4=ME=mixing equipment, a categorical factor with levels {M,D} and x_5=DS addition, also categorical with levels {S,P}. All factors were coded into the (-1,1) scale. The categorical factors were modeled each as a Bernoulli(0.5) random variable.

The resulting Robust Parameter Design problem consists in finding RT, RH, and DT that with high probability gives a drug release stability profile at 6 months of storage that does not decay with respect to the reference profile *regardless* of the mixing equipment and DS addition used. This is an instance of a process that generates profiles that change in time (1 to 6 months, in this particular case). In this example, the rational is that if the latest profile (at stability time = 6 months), corresponding to the oldest stability time of drug, satisfies the product specifications, then it can be assured that the rest of the profiles at earlier stability times, will also satisfy them.

The drug release vs. release time profiles are percentages, and hence, have a form that would not follow a normal distribution, necessary in our approach. We therefore first normalize the data by applying the Probit transformation:

$$Y' = \Phi(Y)^{-1}$$

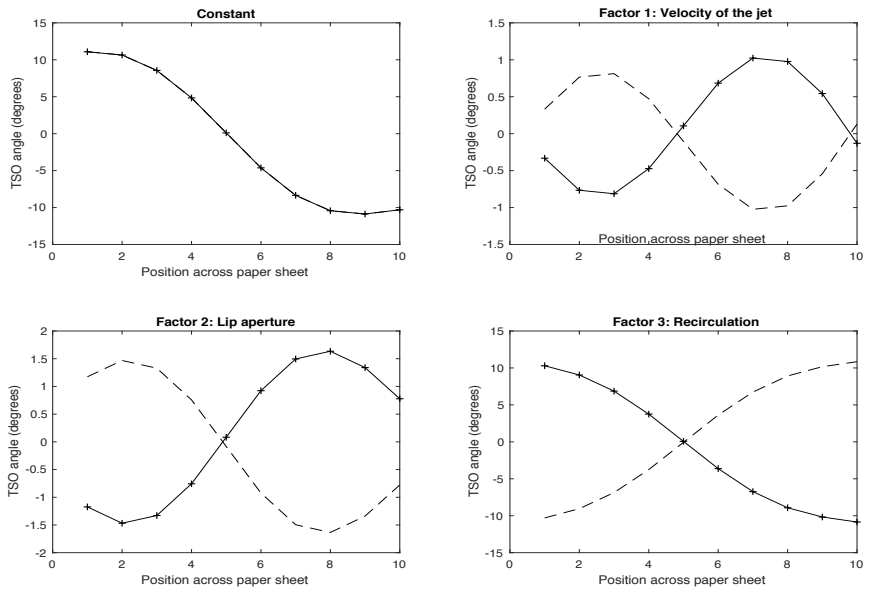

Figure 8: Effect of each of the 3 controllable variables on the TSO response across the paper sheet. The difference between the two lines shows the effect on the curve or profile response that the factor has, as it changes from a low to a high setting. Lines with '+' correspond to the average response when the factor was equal to its highest experimental setting, dashed lines are the average TSO response when the factor was equal to its lowest experimental setting.

where $\Phi(\cdot)$ is the cumulative distribution function of a standard normal random variable. The response Y is the drug release percentage curve at the last observed stability time (6 months). The experimental ran was a $N = 25$ D-optimal design which allows us to estimate all main effects and 2-factor interactions plus quadratic terms in the non-categorical factors (18 parameters). The observed curves at each of the 25 experimental runs are shown in figure 11, together with the predicted profiles. The optimal settings found (assuming both noise factors are randomly varying as Bernoulli(0.50) random variables) are shown in table 7.

	RT=x_1	RH=x_2	DT=x_3
x^* (coded)	0.855	0.848	0.841
x^* (uncoded)	24.42	59.57	58.48

Table 7: Optimal settings for Drug Release experiment.

The optimal profile at the $\mathbf{x}_c^*$ settings is shown in figure 12 together with 95% prediction intervals. The estimated maximum joint probability of having a future curve completely inside the bounds when running the process at settings $\mathbf{x}_c^*$ is 0.4060. The joint probability is low because of the difficulty of maintaining the response within bounds at the highest release times, when it is more variable. Finally, figure 13 shows the main effect plots of factors RT, RH, DT, ME, and DS. The latter two are categorical factors. From the plots, to keep the release curve as high as possible, especially with a high increase initially, all

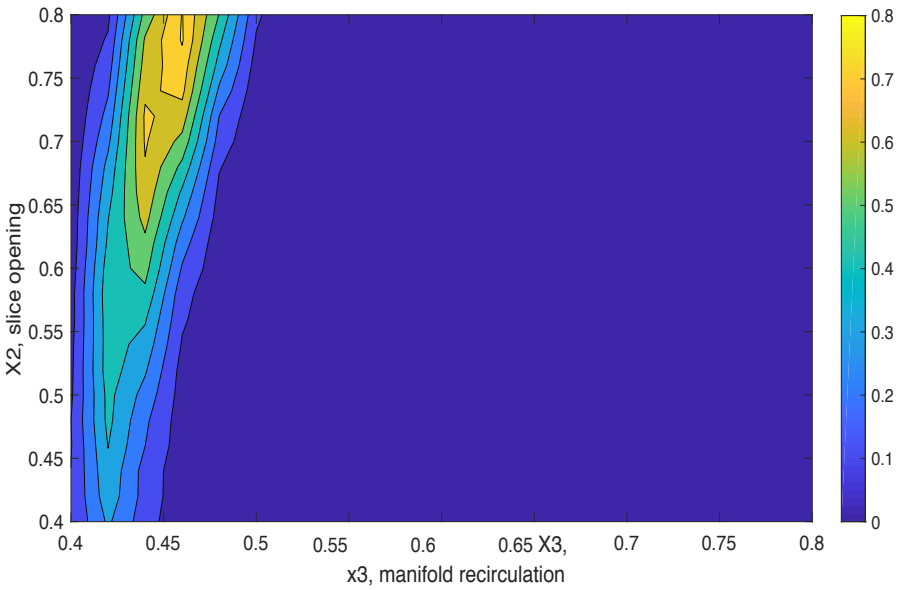

Figure 9: Contour plot of the predictive probability $p(\mathbf{Y} \in A|\text{data}, \mathbf{x})$ for $x_1 = 847.4$ while varying x_2 and x_3 within their feasible region (40 % to 80 %). A given contour can be used as the boundary defining the design space for this particular process.

3 controllable factors, RT, RH, and DT should be run at their high settings. This agrees with the optimal solution, obtained via numerical optimization, displayed in Table 7. A confirmation run was made at these settings. Figure 14 shows the actual drug release profiles at different stability times.

Finally, a simple type of optimization can be done over the categorical factors if desired. This can be achieved by systematically varying the probabilities p_1 and p_2 of the Bernoulli(p_i) variables x_4 and x_5, and solve the associated bayesian RPD problems again. Table 8 shows the results obtained in this analysis. A value of $p_i = 1$ is equivalent to say that the corresponding factor was fixed at its low (-1) setting. While the optimal settings of the controllable factors change little when ME and RS are varied, it is clear that highest probabilities of conformance to specifications can be obtained when these two categorical factors have *opposing* settings, either $(x_1, x_2) = (-1, 1)$ or (1,-1) corresponding to the (ME, DS)=(M,P) or (D,S) settings.

4 Comparing the hierarchical model for profile responses to the probabilistic latent variable models

Profile responses are increasingly common in the process industry [52], and their dimensionality, i.e., the number of observations per profile, J can be high. A frequent approach in these high-dimensional settings is to adopt latent variable approaches, where the goal is to find one or more lower dimensional (latent) spaces of variables that structure and simplify the analysis of the input-output data. In this section, we compare the structure of

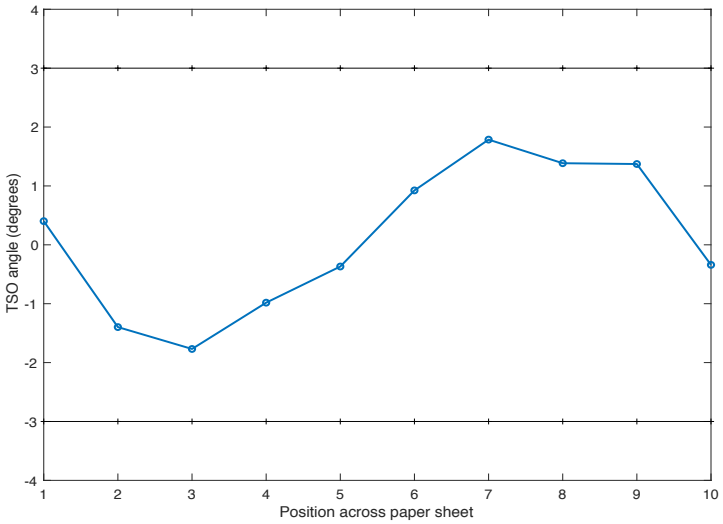

Figure 10: Blue line with dots: TSO profile from confirmation run at $VJ = 847.4$, $A = 79.9$ and $VR = 45.35$. Crossed dark lines: upper and lowed specifications for the TSO profile.

p_1	p_2	$p(\mathbf{Y}(\mathbf{x}_c) \in A \vert \text{data})$	x_1^*=RT	x_2^*=RH	x_3^*=DT	x_4 =ME	x_5 =RS
0.5	0.5	0.406	0.855	0.848	0.841		
1	1	0.416	0.814	0.998	0.999	-1	-1
1	0	0.496	1.000	0.754	0.917	-1	1
0	1	0.505	0.889	0.830	0.999	1	-1
0	0	0.311	0.792	0.811	0.910	1	1

Table 8: Optimal settings for the drug release experiment when noise factors x_4=ME and x_5=RS are assumed categorical with Bernoulli distributions with probabilities p_1 and p_2 respectively. The optimal solutions for controllable factors x_1, x_2, x_3 remain largely the same.

probabilistic latent variable methods with the hierarchical model used for profile responses (9-10).

The latent variable approaches that are most commonly applied in the chemometrics/process analytics communities are Principal Components Regression (PCR), Partial Least Squares (PLS) and Canonical Correlation Analysis (CCA). They are classically presented as sample estimation procedure without reference to an explicit stochastic model that represents the population from which the $\mathbf{X}$ and $\mathbf{Y}$ data were generated. In PCR one estimates the principal components of $\mathbf{X}$ and uses them as explanatory regressors for $\mathbf{Y}$; PLS finds the latent variables common to $\mathbf{X}$ and $\mathbf{Y}$ from maximizing the covariance between these matrices; and CCA finds the linear combinations maximizing the correlation between $\mathbf{X}$ and $\mathbf{Y}$ (instead of the covariance, as in PLS) [55, 26, 28, 6, 30, 16]. Until recently, this has been the prevailing perspective from the chemometrics/process analytics communities.

The Machine Learning community looks at these methods from a probabilistic perspective, where a stochastic model explains how the data were generated, and based upon which

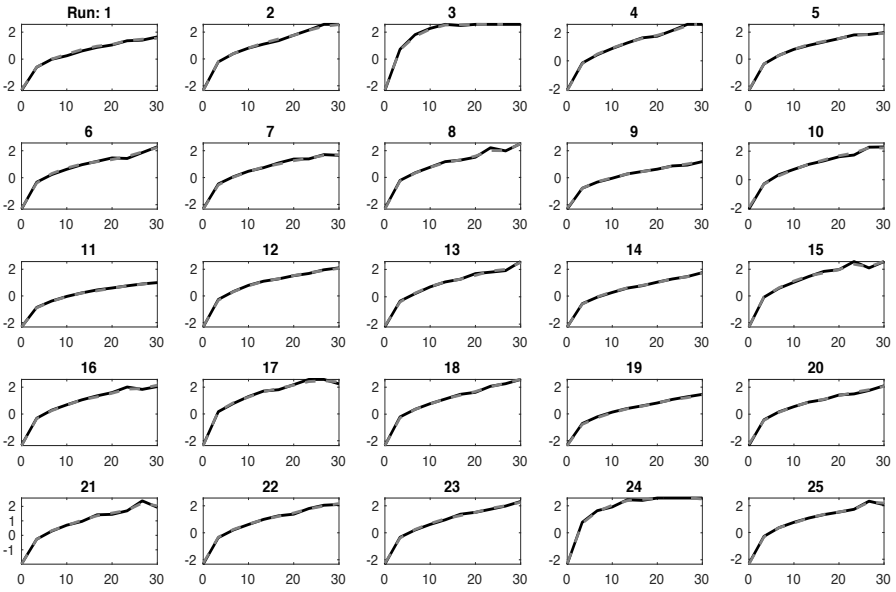

Figure 11: Dark lines: observed 6-month drug release profiles in each of the $N = 25$ experimental runs (Probit transformation applied). Lighter, dashed lines: predicted profiles (mean of the posterior distribution). The predictions are so close to the actual they are hard to see.

maximum likelihood estimators of the different latent variable can be obtained, from which bayesian approaches can be developed. Starting with the work by Tipping and Bishop [61] on probabilistic PCA (see Appendix C), various authors have developed probabilistic versions of PCR, PLS, and CCA. Ge [21] has recently reviewed probabilistic latent variable models and their application in chemometrics and in general, in the process industries. We refer to Ge's paper and to [5] and [36] for fuller accounts, and only summarize next these probabilistic models given our goal of contrasting their structure with the hierarchical mixed effects model we used for profile response experiments, highlighting their different structure.

Probabilistic Principal Components Regression. In PCR (model *a)* in figure 15), it is assumed that both the response space $\mathbf{Y}$ and the controllable factor space $\mathbf{X}$ share a common latent space Z (where X, Y, and Z are vectors spaces where the corresponding variables lie). The model (assumed centered input and output data, so their means are all zero) is then

$$\begin{aligned} \mathbf{y}|\mathbf{z} &= \mathbf{W}\mathbf{z} + \boldsymbol{\varepsilon}_1, \quad \varepsilon_1 \sim N(\mathbf{0}, \boldsymbol{\Sigma}) \quad (\boldsymbol{\Sigma} \text{ diagonal}) \\ \mathbf{x}|\mathbf{z} &= \mathbf{A}\mathbf{z} + +\boldsymbol{\varepsilon}_2, \quad \varepsilon_2 \sim N(\mathbf{0}, \sigma^2 I) \end{aligned}$$

where $\mathbf{y} \in Y, \mathbf{x} \in X$ and the latent variables $\mathbf{z} \in Z$ are of much lower dimension than the $\dim(\mathbf{x})$ or $\dim(\mathbf{y})$. From these assumptions, the relation of most interest for process optimization, $\mathbf{y}|\mathbf{x}$, can be obtained [36]. Chen et al. [7] give an excellent presentation of Markov Chain Monte Carlo methods applied to the PCR model, with practical application in chemical processes.

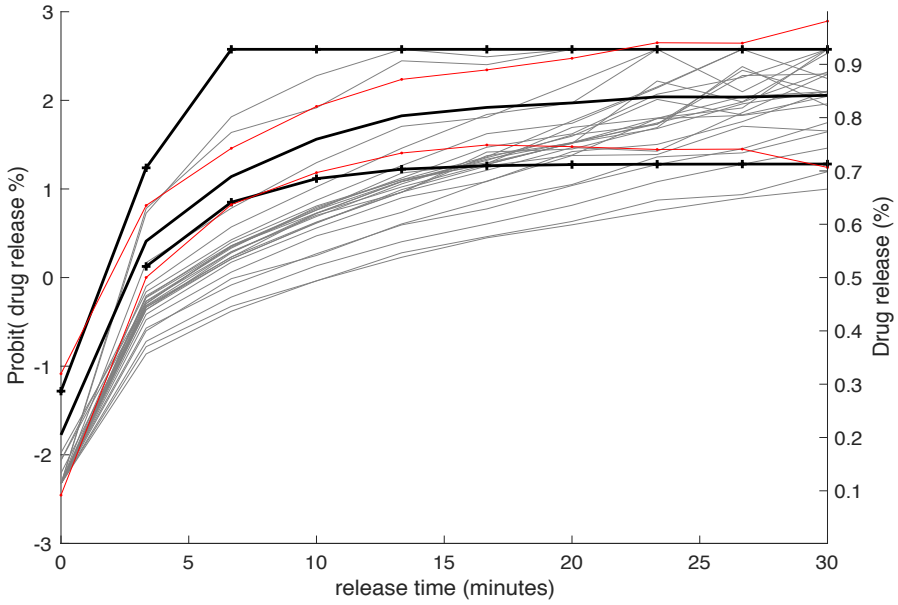

Figure 12: Light lines: observed drug release profiles after a 6 month stability period. A probit ($Y' = \Phi(Y)^{-1}$) transformation was applied to the drug release percentage response before analysis. Actual drug release % can be read from the right hand axis. Dark line: optimal predicted drug release profile at 6 months, red lines: the 5 and 95% predicted percentiles of the optimal drug release posterior distribution, crossed dark lines: upper and lowed specifications for the 6-month drug release profile.

Probabilistic Partial Least Squares. In PLS, in addition of a latent space Z^c that is common to both Y and X spaces, there is a unique latent space for the input space, Z^x (see model b) in figure 15). The model (assuming centered input and output data) is:

$$\begin{aligned} \mathbf{y}|\mathbf{z} &= \mathbf{W}\mathbf{z}^c + \boldsymbol{\varepsilon}_1, \quad \boldsymbol{\varepsilon}_1 \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \\ \mathbf{x}|\mathbf{z} &= \mathbf{A}\mathbf{z}^c + \mathbf{B}\mathbf{z}^x + \boldsymbol{\varepsilon}_2, \quad \boldsymbol{\varepsilon}_2 \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \end{aligned}$$

Probabilistic Canonical correlation analysis. In addition to the structure in PCA, in CCA there is also a unique latent space Z^y for Y (model c) in figure 15). The model, under centered input and output data is:

$$\begin{aligned} \mathbf{y}|\mathbf{z} &= \mathbf{W}\mathbf{z}^c + \mathbf{C}\mathbf{z}^y + \boldsymbol{\varepsilon}_1, \quad \boldsymbol{\varepsilon}_1 \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \\ \mathbf{x}|\mathbf{z} &= \mathbf{A}\mathbf{z}^c + \mathbf{B}\mathbf{z}^x + \boldsymbol{\varepsilon}_2, \quad \boldsymbol{\varepsilon}_2 \sim N(\mathbf{0}, \sigma^2\mathbf{I}) \end{aligned}$$

In both PLS and CCA, the induced distribution $\mathbf{y}|\mathbf{x}$, can also be obtained but numerical integration is necessary [36].

In contrast with the previous probabilistic latent models, in the process optimization setting we discuss in this paper, the X space is usually well defined and *not* high dimensional, as it comes from an experimental design where the number of controllable factors is usually not large. Therefore, there is no need to find a latent structure in it, as there are no underlying unobserved variables generating the observed variability in X, on the

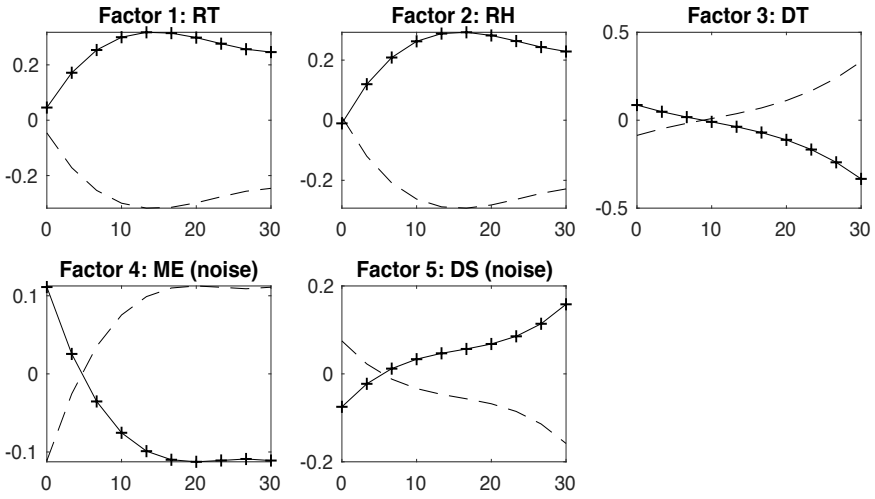

Figure 13: Effect of each of the 3 controllable variables of the drug release profiles at 6 month stability time. The difference between the two lines shows the effect on the curve or profile response that the factor has, as it changes from a low to a high setting. Lines with '+' correspond to the average response when the factor was equal to its highest experimental setting, dashed lines are the average response when the factor was equal to its lowest experimental setting. Response (% drug release) shown after Probit transformation.

contrary, such variation is induced by the measured factors. For the same reason, *all of the latent space models above, which assume a common latent structure in the X and Y spaces are unnecessary and do not correspond to the data structure of an industrial experiment.* Dimensionality reduction may be needed in the response space only, where a latent variable structure may exist for which a lower dimensional space can be estimated. For instance, the multivariate linear regression model discussed earlier is applicable when J is not too large compared to N, because otherwise this would require weighting priors that are harder to justify. If the number of responses J is large, a principal component analysis can be conducted on the Y space prior to the analysis, and then a multivariate regression model is built from the X space to the Y latent variable space. The appropriate data structure then goes from X space to Z space to Y space (see figure 16).

The latent space structure depicted in figure 16 is made explicit in the hierarchical model we presented for profile responses, model (9-10). Note how the two-stage hierarchical model can be written as:

$$\begin{aligned} \mathbf{y}|\mathbf{z} &= \mathbf{W}\mathbf{z} + \boldsymbol{\varepsilon}_1, \quad \boldsymbol{\varepsilon}_1 \sim N(0, \Sigma) \quad (\Sigma \text{ diagonal}) \\ \mathbf{z}|\mathbf{x} &= \mathbf{B}\mathbf{x} + +\boldsymbol{\varepsilon}_2, \quad \boldsymbol{\varepsilon}_2 \sim N(0, \sigma^2 I) \end{aligned}$$

by considering the model parameters in stage 1, $\boldsymbol{\theta}$, the latent "features" of the profile or curve response ($\mathbf{z}$). Thus, the hierarchical mixed effects model implements a feature-based dimensionality reduction in the data. It can be used either for profile/curve response systems (where the main curve features are preselected by the user) or for high-dimensional Y-data via a previously conducted principal component analysis that provides the "design" matrix $\mathbf{W}$ above.

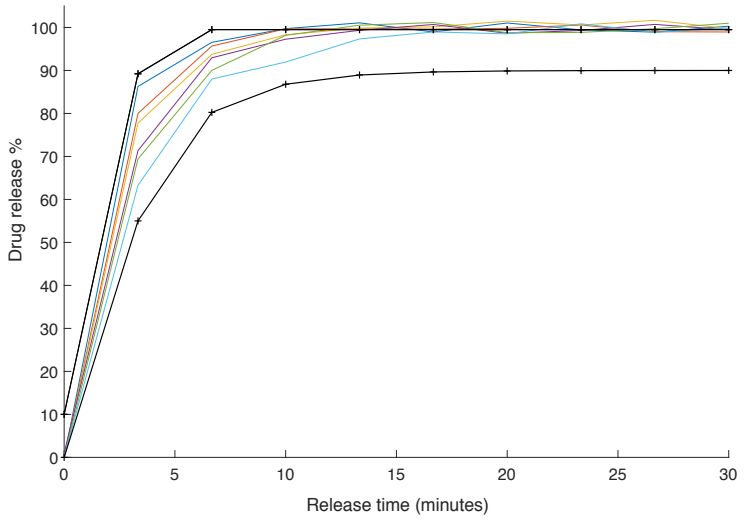

Figure 14: Light lines: observed drug release profiles at stability times 1, 2, 3, 4, 5, and 6 months from confirmation run at $RT = 24.4$, $A = 59.5$ and $VR = 58.4$ ($x_1 = 0.855, x_2 = 0.848, x_3 = 0.841$). Categorical factors were randomly varied as Bernoulli(0.5) random variables, thus the solution is robust with respect to either setting of ME and DS. Crossed dark lines: upper and lowed specifications for the TSO profile.

5 Conclusions

We have provided a review and some extensions of bayesian predictive optimization methods, with application to the process industries. The methods are based on predictors and response data from designed experiments, where the response is in the form of either a vector of correlated responses, possibly of high dimensionality, or a profile. Bayesian predictive methods are becoming increasingly relevant in practice not only for process optimization but also in QbD applications where it is of interest to find a design space where the future performance of the process is guaranteed with certain probability (as in pharmaceutical applications, see [43, 58]). The bayesian predictive approach to process optimization provides a probability (or reliability measure) of satisfying the process goals at the optimal settings $\mathbf{x}_c^*$. Such probability statements are not possible to obtain with classical statistics. A frequentist "design space" can be obtain classically by bootstrapping, a possibility not reviewed here. The bootstrapping approach, however, can only provide confidence regions on the optimal operating conditions $\mathbf{x}_c^*$ and it is not possible to provide *probability statements* about whether the process will fall or not in the given region if in the future it is run at operating conditions $\mathbf{x}_c^*$, something easily done under the bayesian framework.

Extensions to the original bayesian process optimization methodology, based on regression models in [41] were presented. These included the use of a hierarchical mixed effects model that can be used either when the number of responses is too large for a noninformative multivariate regression analysis, in such a way that a preliminary probabilistic PCA provides the "features" to be modeled in a second stage model as a function of controllable factors, or to model profile responses, where a first stage model defines a parametric model expressing the shape of the curve that has been observed, and a second stage models the re-

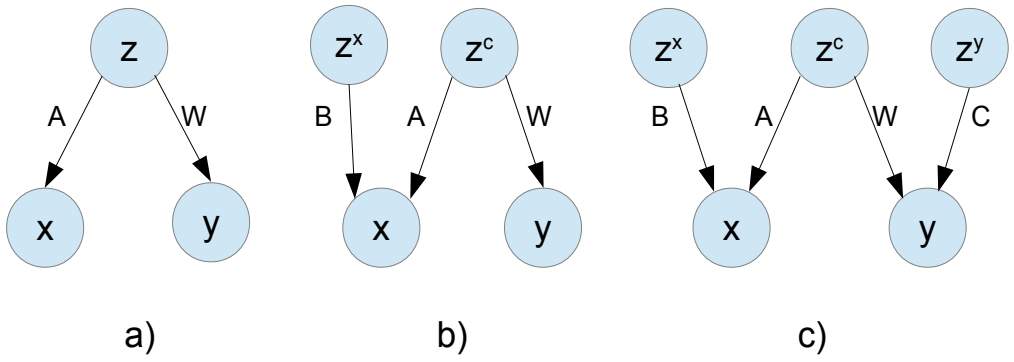

Figure 15: Other alternative model structures with latent variables. a) Principal components regression. b) Partial least squares; c) Canonical correlation analysis (adapted from a figure by [36]). In situations where the input data ($\mathbf{x} \in X$) comes from a designed experiment these structures are not appropriate, as there is no latent space on the X space.

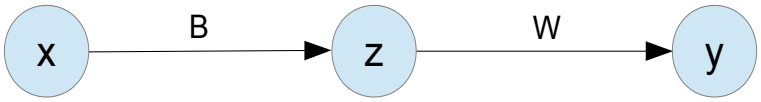

Figure 16: The model structure needed if a high dimensionality response is considered in an experimental optimization situation. A latent structure in the Y space only may be necessary if Y is very high dimensional and condition (2) does not hold.

lation of these parameters with the controllable factors. In either case, a complete bayesian solution is available. The probabilistic latent variable models were shown to have a structure that does not correspond to that of data from a process optimization experiment. We also discussed the case noise factors are present and a "Robust Parameter Design" optimization is desired, and an extension to the case when noise factors are categorical was discussed and illustrated. The MATLAB code provided (see supplementary materials) can be used to replicate the analyses in the case studies covered in this paper, and it is hoped it will facilitate the comprehension of the paper conceptual contents, as well as their application and future adoption.

Appendix A. Calculating the probability of conformance to specifications

The following Monte Carlo procedure can be used to compute $p(\mathbf{x_c})_{\text{RPD}}$, the quantity maximized in problem (7) for the single stage, multiple response approach (see [35] for more details).

1. Set $c = 0$.

2. Simulate a value of the noise factors, $\mathbf{x}_n$, from their (assumed known) distribution. The vector $\mathbf{x}_c$ is given and contains the values of the controllable factors. Make $\mathbf{x} = (\mathbf{x}_c, \mathbf{x}_n)$.

3. Simulate $\mathbf{y}|\mathbf{x}, data \sim \mathbf{T}_J^{\nu}\left(\mathbf{B}'\mathbf{x}, \frac{\nu}{\nu-2}\mathbf{H}^{-1}\right)$

4. If $\mathbf{y} \in A$, make $c \leftarrow c + 1$. Go to step 2 and repeat m times.

5. Return $\widehat{p}(\mathbf{x_c})_{\mathrm{RPD}} = c/m$.

For the hierarchical mixed model, step 3 is substituted by sampling from the simulated Markov Chain obtained via Gibbs sampling (after the burn-in period) and substituting these simulated parameters in the likelihood, i.e., values of $\mathbf{y}$ are obtained from the predictive density by composition (see Appendix B). As reported by Miro et al. [35], it is more effective to generate the required m noise factors once and use a common random numbers strategy in the optimization whenever $p(\mathbf{x_c})_{\mathrm{RPD}}$ needs to be evaluated within the optimization process. Optimizing this function is a hard nonlinear problem, and Miro et al. report how it is useful to start the optimizer from a point likely to contain enough "area" of the predictive density.

Appendix B. Gibbs sampling for estimating the linear mixed model

Lange et al. [33] give the full conditionals for the parameters $(\boldsymbol{\beta}, \{\mathbf{w}_i\}, \{\sigma_i^2\}, \boldsymbol{\Sigma}_w)$ in the linear mixed model (12) where the σ_i^2's allow different variances among the observed profiles (this is sometimes useful in longitudinal analysis models when it is not desired to make inferences on profiles other than the N observed). Chib and Carlin [10] considered the case we consider, where $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_J$, and show how the Lange et al. procedure suffers from slow convergence, and proposed two alternative algorithms, one (their algorithm 2) which is a pure Gibbs sampling approach, and another one (their algorithm 3) which has a Metropolis step (a rejection sampling step). Since the convergence properties of these two algorithms appear to be about the same, in particular for the $\boldsymbol{\beta}$ parameters, we choose Chib and Carlin's algorithm 2 (after correcting some errors in the full conditionals in their paper) since it does not require a Metropolis step (i.e., all full conditionals are known distributions). We also provide the full conditionals of the other parameters since they were not explicitly given by these authors.

The priors we use were:

$\boldsymbol{\beta} \sim N_{pq}(\boldsymbol{\beta}_0, \mathbf{B}_0)$, with $\boldsymbol{\beta}_0 = \mathbf{0}$ and $\mathbf{B}_0 = 1000\mathbf{I}$ (noninformative for $\boldsymbol{\beta}$);
$\sigma^2 \sim IG(\lambda_1, \lambda_2)$ (inverse-gamma distribution), with $\lambda_1 = 0.001$ and $\lambda_2 = 5$ (this gives $E(\sigma^2) = 2$ and $\sqrt{\mathrm{Var}(\sigma^2)} = 63$, relatively non-informative);
$\boldsymbol{\Sigma}_w^{-1} \sim \mathrm{Wishart}(\nu_0^{-1}\mathbf{R}_0, \nu_0)$ with $\nu_0 = p$ (most non-informative choice) and $\mathbf{R}_0 = \mathbf{I}_p$ (gives $E(\boldsymbol{\Sigma}_w) = \mathbf{I}_p$ with as large variance as possible);

The Gibbs sampling scheme is:

1. Sample $\boldsymbol{\beta}$ from $\boldsymbol{\beta}|\mathbf{Y}, \sigma^2, \boldsymbol{\Sigma}_w$:

$$N\left(\left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{V}^{-1} \boldsymbol{X}_i + \mathbf{B}_0^{-1}\right)^{-1} \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{V}^{-1} \mathbf{y}_i + \mathbf{B}_0^{-1} \boldsymbol{\beta}_0\right), \left(\sum_{i=1}^{N} \boldsymbol{X}_i' \boldsymbol{V}^{-1} \boldsymbol{X}_i + \mathbf{B}_0^{-1}\right)^{-1}\right)$$

   where $\boldsymbol{V} = \mathbf{S}^* \boldsymbol{\Sigma}_w \mathbf{S}^{*'} + \sigma^2 \mathbf{I}$ and $\boldsymbol{X}_i = \mathbf{x}_i^{(m)'} \otimes \mathbf{S}^*$.
2. Sample the random effects $\mathbf{w}_i$, $i = 1, ..., N$ from $\{\mathbf{w}_i\}|\mathbf{Y}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\Sigma}_w$:

$$N\left(\left(\frac{\mathbf{S}^{*'}\mathbf{S}}{\sigma^2} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1} \frac{\mathbf{S}^{*'} \boldsymbol{R}_i^{(w)}}{\sigma^2}, \left(\frac{\mathbf{S}^{*'}\mathbf{S}}{\sigma^2} + \boldsymbol{\Sigma}_w^{-1}\right)^{-1}\right)$$

   where $\boldsymbol{R}_i^{(w)} = \mathbf{y}_i - \boldsymbol{X}_i \boldsymbol{\beta}$.
3. Sample $\boldsymbol{\Sigma}_w^{-1}$ from $\boldsymbol{\Sigma}_w^{-1}|\{\mathbf{w}_i\}$:

$$\text{Wishart}\left(\left(\sum_{i=1}^{N} \mathbf{w}_i \mathbf{w}_i' + \nu_0 \mathbf{R}_0^{-1}\right)^{-1}, N + \nu_0\right)$$

4. Sample σ^{-2} from $\sigma^{-2}|\mathbf{Y}, \boldsymbol{\beta}, \{\mathbf{w}_i\}$:

$$\text{Gamma}\left(\lambda_1 + \frac{JN}{2}, \left(\frac{1}{\lambda_2} + \frac{1}{2}\left(\mathbf{R}^{(\sigma^2)'} \mathbf{R}^{(\sigma^2)}\right)\right)^{-1}\right)$$

   where $\mathbf{R}^{(\sigma^2)} = \mathcal{Y} - \mathcal{X}\boldsymbol{\beta} - \text{vec}(\mathbf{S}^* \mathbf{w}_1, ..., \mathbf{S}^* \mathbf{w}_N)$ (a $NJ \times 1$ vector).

The MCMC sampling of the mixed effects model parameters needs not be conducted within the optimization routine necessary to solve (7), otherwise this would imply a tremendous computational burden. The reason for this is given by the model written as in (11). Given realizations of the posterior of $\boldsymbol{\Theta} = (\boldsymbol{\beta}, \{\mathbf{w}_i\}\boldsymbol{\Sigma}_w, \sigma^2)$, $p(\boldsymbol{\Theta}|data)$, we can simulate draws of the posterior predictive density by composition (Gelman et al., 2004):

$$\begin{aligned} p(\mathbf{y}|data, \mathbf{x}) &= \int p(\mathbf{y}, \boldsymbol{\Theta}|data, \mathbf{x}) d\boldsymbol{\Theta} \\ &= \int p(\mathbf{y}|data, \boldsymbol{\Theta}, \mathbf{x}) p(\boldsymbol{\Theta}|data) d\boldsymbol{\Theta} \end{aligned}$$

Thus, we conduct a simulation of the Gibbs sampling chain until convergence, approximating in this way $p(\boldsymbol{\Theta}|data)$ and simply sample from it. We then substitute the sampled values into the marginal likelihood $p(\mathbf{y}|data, \boldsymbol{\Theta}, \mathbf{x})$ whenever a $\mathbf{y}|data, \mathbf{x}$ vector is needed in the optimization routine (this replaces step 3 in Appendix A). Hence, the MCMC computations are only run once, before performing the optimization. For more information on MCMC methods, see Gelman et al. [23] and for a more concise introduction, see Colosimo and Del Castillo [11].

Appendix C. Probabilistic Principal Components Analysis

In order to present the probabilistic version of PCA of Tipping and Bishop [61], we first contrast the closely related Maximum Likelihood PCA (MLPCA) models in the Chemometrics literature (e.g., as in [64]) and the Factor Analysis (FA) model in the Machine Learning and Statistics literature (e.g., as in [5, 24])[1]. We then discuss probabilistic PCA, as it was used in the estimation of matrix $\mathbf{S}^*$ in the linear mixed effects model (12).

Both MLPCA and FA models are based on the description:

$$\mathbf{Y} = \mathbf{W}\mathbf{z} + \boldsymbol{\mu} + \boldsymbol{\varepsilon}$$

where $\mathbf{Y}$ is a p-dimensional observed vector and $\mathbf{z}$ is an unobserved (and therefore, hypothesized) k dimensional vector, with k typically much smaller than p. W is called the "loadings" matrix and the entries in the $\mathbf{z}$ vector are called the latent or "factor" variables. Both MLPCA and Factor Analysis assume:

$$\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}).$$

If we redefine $\mathbf{Y} - \boldsymbol{\mu}$ to be observed data (i.e., if we center the data with $\hat{\boldsymbol{\mu}} = \overline{\mathbf{Y}}$), the parameter $\boldsymbol{\mu}$ can be neglected.

MLPCA model characteristics.- The goal of MLPCA according to [64], is to best estimate $\mathbf{z}$ given the observed random vectors $\mathbf{Y}$, an estimate of $\mathbf{W}$ and a *known* covariance matrix $\boldsymbol{\Sigma}$. In MLPCA it is assumed $\mathbf{z}$ is an error-free constant, so the model is a linear regression model, and the optimal solution is given by the generalized least squares estimator:

$$\hat{\mathbf{z}} = (\mathbf{W}'\boldsymbol{\Sigma}^{-1}\mathbf{W})^{-1}\mathbf{W}'\boldsymbol{\Sigma}^{-1}\mathbf{Y}$$

which yields predictions of the responses $\hat{\mathbf{Y}} = \mathbf{W}\hat{\mathbf{z}} + \hat{\boldsymbol{\mu}}$. This prediction equation is used to minimize the "reconstruction error"

$$\mathbf{S}_\Sigma = \sum_{i=1}^{N} (\hat{\mathbf{Y}} - \mathbf{Y}_i)'\boldsymbol{\Sigma}^{-1}(\hat{\mathbf{Y}} - \mathbf{Y}_i)$$

with respect to $\mathbf{W}$. The MLPCA literature also discusses the case where $\boldsymbol{\Sigma}$ varies with each observation, so the the covariance between the same elements of different $\mathbf{Y}_i$'s can differ with the observation number i. Covariances between elements in each $\mathbf{Y}_i$ are modeled via the entries of $\boldsymbol{\Sigma}_i$, which is a dense (not diagonal) matrix. Note that in processes where a very large number of variables p are measured the number of elements in $\boldsymbol{\Sigma}_i$ is high ($p(p+1)/2$ distinct entries). These covariances are assumed known via information available for the measurement noise, although [64] mention that in practice they need to

[1]The Machine Learning and Statistics literature actually calls "PCA" the model where $\Sigma = \sigma^2 I$ while it calls the model a FA model when Σ is diagonal but has different entries. Thus PCA in Chemical Engineering is closer to FA in Statistics and Machine Learning.

be estimated[2]. The objective of MLPCA in Chemometrics then is to find relationships between the measured variables.

FA model characteristics.- The FA model in the Machine Learning/Statistics literature, rather than considering the latent variables fixed constants, assumes instead they are random:

$$\mathbf{z} \sim N(\mathbf{0}, \mathbf{\Sigma}_z).$$

This is a "random effects" model with two sources of variability affecting the observations. The term containing the latent variables $\mathbf{z}$ models correlations between the elements of $\mathbf{Y}$, while the error variables $\boldsymbol{\varepsilon}$ account for measurement error in each of these elements. In the Chemometrics field, Reis and Saraiva [51] use random latent variables (in contrast to Wenzell [64]), calling the model an heteroscedastic latent model, and use it for Statistical Process Control.

The goal when using the FA model is to estimate $\mathbf{W}$ *and* $\mathbf{\Sigma}$ under the additional assumption that $\mathbf{\Sigma}_z = \mathbf{I}_k$ (the k dimensional identity matrix) to best model the correlation in the entries of the observed $\mathbf{Y}$ vectors. This additional assumption does not lose generality because any correlation between the elements of $\mathbf{Y}$ can still be modeled given that

$$Cov(\mathbf{Y}) = \mathbf{W}\mathbf{W}' + \mathbf{\Sigma}. \tag{13}$$

A particular case: Probabilistic PCA.- If $\mathbf{\Sigma} = \sigma^2\mathbf{I}$ (same measurement error variance for all elements in $\mathbf{Y}$) the FA model is called a **Probabilistic PCA** model (PPCA). The maximum likelihood estimates of $\mathbf{W}$ and σ^2 were shown by [61] to be:

$$\hat{\mathbf{W}} = \mathbf{V}(\mathbf{L} - \sigma^2\mathbf{I})^{-1/2}\mathbf{R}$$

where V is a $p \times k$ matrix with columns equal to the k eigenvectors associated with the top k eigenvalues of

$$\mathbf{S} = \sum_{i=1}^{N}(\hat{\mathbf{Y}} - \mathbf{Y}_i)(\hat{\mathbf{Y}} - \mathbf{Y}_i)'$$

$\mathbf{L}$ is a diagonal matrix with the k largest eigenvalues, and $\mathbf{R}$ is an arbitrary rotation matrix, which can be set equal to $\mathbf{I}$. The MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{p-k}\sum_{i=k+1}^{p}\lambda_i \tag{14}$$

(the variance associated with the discarded dimensions). The inverse mapping in PPCA, giving the latent vector associated with a given observation $\mathbf{Y}$, is the posterior distribution of $\mathbf{z}$, which is

$$\mathbf{z}|\mathbf{Y} \sim N((\mathbf{W}'\mathbf{W} + \sigma^2\mathbf{I})^{-1}\mathbf{W}'\mathbf{Y},\ \sigma^2(\mathbf{W}'\mathbf{W} + \sigma^2\mathbf{I})^{-1})$$

As it can be seen, the posterior mean is a linear projection that can be interpreted as "ridge" regression [25].

[2]A computational consideration for very large p is that the dense $p \times p$ matrix Σ needs to be inverted to find $\hat{\mathbf{z}}$, which is an $O(p^3)$ operation.

Supplementary materials

All datafiles and MATLAB code that implements the methods discussed here, including MCMC estimation and numerical optimization are provided. Running the script `scriptExamples.m` reproduces all examples in the paper by calling the appropriate functions. In addition, function `Find_PCA_Dim.m` is provided, which implements the exact ML solution for Probabilistic PCA. It also finds the best latent space dimension using a bayesian algorithm developed by Minka [34].

References

[1] ASI. *Robust Designs Using Taguchi Methods*. American Supplier Institute (ASI) Press, Livonia, MI., 1998.

[2] Bhavik R Bakshi, Mohamed N Nounou, Prem K Goel, and Xiaotong Shen. Multiscale bayesian rectification of data from linear steady-state and dynamic systems without accurate models. *Industrial and Engineering Chemistry Research*, 40(1):261–274, 2001.

[3] Yaakov Bar-Shalom and Leon Campo. The Effect of the Common Process Noise on the Two-Sensor Fused-Track Covariance. *IEEE Transactions on Aerospace and Electronic Systems*, 22(6):803–805, November 1986.

[4] Dean Billheimer. Predictive inference and scientific reproducibility. *The American Statistician*, 73(sup1):291–295, 2019.

[5] Christopher M Bishop. *Pattern recognition and machine learning*. Springer Science+ Business Media, 2006.

[6] A.J. Burnham, John F. MacGregor, and R. Viveros. Latent variable multivariate regression modeling. *Chemometrics and Intelligent Laboratory Systems*, 48(2):167–180, 1999.

[7] Hongshu Chen, Bhavik R Bakshi, and Prem K Goel. Bayesian latent variable regression via gibbs sampling: methodology and practical aspects. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 21(12):578–591, 2007.

[8] Hongshu Chen, Bhavik R Bakshi, and Prem K Goel. Toward bayesian chemometrics—a tutorial on some recent advances. *Analytica chimica acta*, 602(1):1–16, 2007.

[9] Wen-shiang Chen, Bhavik R Bakshi, Prem K Goel, and Sridhar Ungarala. Bayesian estimation of unconstrained nonlinear dynamic systems. *IFAC Proceedings Volumes*, 37(1):263–268, 2004.

[10] S. Chib and B.P. Carlin. On mcmc sampling in hierarchical longitudinal models. *Statistics and Computing*, 9:17–26, 1999.

[11] B.M. Colosimo and E. Del Castillo. Modern numerical methods in bayesian computation. In *Bayesian Process Monitoring, Control and Optimization*, NY, 2007. CRC/Taylor and Francis.

[12] E. Del Castillo. *Process Optimization, a Statistical Approach*. Springer, 2007.

[13] E. Del Castillo and B.M. Colosimo. An introduction to bayesian inference in process monitoring, control, and optimization. In *Bayesian Process Monitoring, Control and Optimization*, NY, 2007. CRC/Taylor and Francis.

[14] Enrique Del Castillo, Bianca M Colosimo, and Hussam Alshraideh. Bayesian modeling and optimization of functional responses affected by noise factors. *Journal of Quality Technology*, 44(2):117–135, 2012.

[15] George Derringer and Ronald Suich. Simultaneous optimization of several response variables. *Journal of Quality Technology*, 12(4):214–219, 1980.

[16] William R. Dillon and Matthew Goldstein. *Multivatiate Analysis - methods and applications*. Wiley, 1984.

[17] Alireza Fatehi and Biao Huang. Kalman filtering approach to multi-rate information fusion in the presence of irregular sampling rate and variable measurement delay. *Journal of Process Control*, 53:15–25, May 2017.

[18] FDA. Guidance document: Q11 development and manufacture of drug substances, 2012.

[19] G.M. Fitzmaurice, N.M. Laird, and J.H. Ware. *Applied Longitudinal Analysis*. John Wiley & Sons, Hoboken, NJ, 2004.

[20] J.B. Gao and C.J. Harris. Some remarks on Kalman filters for the multisensor fusion. *Information Fusion*, 3(3):191–201, September 2002.

[21] Zhiqiang Ge. Process data analytics via probabilistic latent variable models: a tutorial review. *I&EC Research*, 57:12646–12661, 2018.

[22] Zhiqiang Ge, Muguang Zhang, and Zhihuan Song. Nonlinear process monitoring based on linear subspace and bayesian inference. *Journal of Process Control*, 20(5):676–688, 2010.

[23] Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. Chapman and Hall/CRC, 2013.

[24] Zoubin Ghahramani, Geoffrey E Hinton, et al. The EM algorithm for mixtures of factor analyzers. Technical report, Technical Report CRG-TR-96-1, University of Toronto, 1996.

[25] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, 2015.

[26] Agnar Höskuldsson. PLS regression methods. *Journal of Chemometrics*, 2(3):211–228, 1988.

[27] Shuo-Huan Hsu, Stephen D Stamatis, James M Caruthers, W N Delgass, V Venkatasubramanian, Gary E Blau, M Lasinski, and S Orcun. Bayesian framework for building kinetic models of catalytic systems. *Industrial and Engineering Chemistry Research*, 48(10):4768–4790, 2009.

[28] Ulf G. Indahl, H. Liland, Kristian, and Tormod Næs. Canonical partial least squares—a unified PLS approach to classification and regression problems. *Journal of Chemometrics*, 23(9):495–504, 2009.

[29] Qingchao Jian, Biao Huang, and Xuefeng Yan. GMM and optimal principal components-based bayesian method for multimode fault diagnosis. *Computers and Chemical Engineering*, 54:338–349, 2008.

[30] R.A. Johnson and D.W. Wichern. *Applied Multivariate Statistical Analysis*. Prentice Hall, 2002.

[31] Hiromasa Kaneko and Kimito Funatsu. Adaptive soft sensor based on online support vector regression and bayesian ensemble learning for various states in chemical plants. *Chemometrics and Intelligent Laboratory Systems*, 20:57–66, 2014.

[32] N.M. Laird and J.H. Ware. Random effects models for longitudinal data. *Biometrics*, 38(4):963–974, 1982.

[33] Nicholas Lange, Bradley P Carlin, and Alan E Gelfand. Hierarchical Bayes models for the progression of HIV infection using longitudinal cd4 t-cell numbers. *Journal of the American Statistical Association*, 87(419):615–626, 1992.

[34] Thomas P Minka. Automatic choice of dimensionality for PCA. In *Advances in neural information processing systems*, pages 598–604, 2001.

[35] G. Miro, E. Del Castillo, and J.J. Peterson. A bayesian approach for multiple response surface optimization in the presence of noise variables. *Journal of Applied Statistics*, 31(3):251–270, 2004.

[36] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[37] R.H. Myers, D.C. Montgomery, and C. Anderson-Cook. *Response Surface Methodology, 3rd. ed.* Wiley, NY, 2009.

[38] Szu Hui Ng. A bayesian model-averaging approach for multiple-response optimization. *Journal of Quality Technology*, 42(1):52–68, 2010.

[39] Mohamed N Nounou, Bhavik R Bakshi, Prem K Goel, and Xiaotong Shen. Bayesian principal component analysis. *Journal of Chemometrics*, 16(9):576–595, 2002.

[40] Mohamed N Nounou, Bhavik R Bakshi, Prem K Goel, and Xiaotong Shen. Process modeling by bayesian latent variable regression. *AIChE journal*, 48(8):1775–1793, 2002.

[41] J.J. Peterson. A bayesian reliability approach to multiple response surface optimization. *Journal of Quality Technology*, 36(2):139–153, 2004.

[42] J.J. Peterson and E. Del Castillo. A posterior predictive approach to multiple response surface optimization. In *JSM, Salt Lake City, Utah*, 2007.

[43] John J. Peterson and Kevin Lief. The ICH Q8 definition of design space: A comparison of the overlapping means and the bayesian predictive approaches. *Statistics in Biopharmaceutical Research*, 2(2):249–259, 2010.

[44] John J Peterson, Guillermo Miro-Quesada, and Enrique del Castillo. A bayesian reliability approach to multiple response optimization with seemingly unrelated regression models. *Quality Technology & Quantitative Management*, 6(4):353–369, 2009.

[45] Alexey L. Pomerantsev. Successive bayesian estimation of reaction rate constants from spectral data. *Chemometrics and Intelligent Laborarory Systems*, 66(2):127–139, 2003.

[46] S.J. Press. *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Analysis*. R. E. Krieger Pub. Co., Malabar, FL., 1982.

[47] Ramkumar Rajagopal and Enrique Del Castillo. Model-robust process optimization using bayesian model averaging. *Technometrics*, 47(2):152–163, 2005.

[48] Ramkumar Rajagopal, Enrique Del Castillo, and John J Peterson. Model and distribution-robust process optimization with noise factors. *Journal of Quality Technology*, 37(3):210–222, 2005.

[49] Marco S Reis, Ana C Pereira, João M Leça, Pedro M Rodrigues, and José C Marques. Multiresponse and multiobjective latent variable optimization of modern analytical instrumentation for the quantification of chemically related families of compounds: Case study—solid-phase microextraction (spme) applied to the quantification of analytes with impact on wine aroma. *Journal of Chemometrics*, 33(3):e3103, 2019.

[50] Marco S Reis and Pedro M Saraiva. Integration of data uncertainty in linear regression and process optimization. *AIChE journal*, 51(11):3007–3019, 2005.

[51] Marco S Reis and Pedro M Saraiva. Heteroscedastic latent variable modelling with applications to multivariate statistical process control. *Chemometrics and Intelligent Laboratory Systems*, 80(1):57–66, 2006.

[52] Marco S Reis and Pedro M Saraiva. Prediction of profiles in the process industries. *Industrial & Engineering Chemistry Research*, 51(11):4254–4266, 2012.

[53] Sajjad Safari, Faridoon Shabani, and Dan Simon. Multirate multisensor data fusion for linear systems using Kalman filters and a neural network. *Aerospace Science and Technology*, 39:465–471, December 2014.

[54] Geisser Seymour. *Predictive Inference*. Chapman and Hall, New York, 1993.

[55] Ruijie Shi and John F. MacGregor. Modeling of dynamic systems using latent variable and subspace methods. *Journal of Chemometrics*, 14(5-6):423–439, 2000.

[56] Branca MA Silva, Sílvia Vicente, Sofia Cunha, Jorge FJ Coelho, Cláudia Silva, Marco Seabra Reis, and Sérgio Simões. Retrospective quality by design (rqbd) applied to the optimization of orodispersible films. *International journal of pharmaceutics*, 528(1-2):655–663, 2017.

[57] Andrew Smyth and Meiliang Wu. Multi-rate Kalman filtering for the data fusion of displacement and acceleration response measurements in dynamic system monitoring. *Mechanical Systems and Signal Processing*, 21(2):706–723, February 2007.

[58] Gregory W. Stockdale and Aili Cheng. Finding design space and a reliable operating region using a multivariate bayesian approach with experimental design. *Quality Technology & Quantitative Management*, 6(4):391–408, 2009.

[59] J.E. Tabora, F. Lora Gonzalez, and J.W. Tom. Bayesian probabilistic modeling in pharmaceutical process development. *AIChE journal*, 65(11), 2019.

[60] G. Taguchi. *System of Experimental Design*. Unipub/Kraus International Publications., NY, 1987.

[61] Michael E Tipping and Christopher M Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(3):611–622, 1999.

[62] O Arda Vanli and Enrique Del Castillo. Bayesian approaches for on-line robust parameter design. *IIE Transactions*, 41(4):359–371, 2009.

[63] James H Ware. Linear models for the analysis of longitudinal studies. *The American Statistician*, 39(2):95–101, 1985.

[64] Peter D Wentzell, Darren T Andrews, David C Hamilton, Klaas Faber, and Bruce R Kowalski. Maximum likelihood principal component analysis. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 11(4):339–366, 1997.

[65] C.F.J Wu and M. Hamada. *Experiments, Planning, Analysis and Optimization, 2nd ed.* Wiley, NY, 2011.

[66] Jie Yu and S. Joe Qin. Multimode process monitoring with bayesian inference-based finite gaussian mixture models. *AIChE Journal*, 54(7):1811–1829, 2008.