Technometrics

A Journal of Statistics for the Physical, Chemical, and Engineering Sciences

Technometrics

ISSN: 0040-1706 (Print) 1537-2723 (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

A Latent Variable Approach to Gaussian Process Modeling with Qualitative and Quantitative Factors

Yichi Zhang, Siyu Tao, Wei Chen & Daniel W. Apley

To cite this article: Yichi Zhang, Siyu Tao, Wei Chen & Daniel W. Apley (2020) A Latent Variable Approach to Gaussian Process Modeling with Qualitative and Quantitative Factors, Technometrics, 62:3, 291-302, DOI: 10.1080/00401706.2019.1638834

To link to this article: https://doi.org/10.1080/00401706.2019.1638834

TECHNOMETRICS
2020, VOL. 62, NO. 3, 291–302
https://doi.org/10.1080/00401706.2019.1638834

Taylor & Francis
Taylor & Francis Group

A Latent Variable Approach to Gaussian Process Modeling with Qualitative and Quantitative Factors

Yichi Zhang[a], Siyu Tao[a], Wei Chen[a], and Daniel W. Apley[b]

[a]Department of Mechanical Engineering, Northwestern University, Evanston, IL; [b]Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL

ABSTRACT

Computer simulations often involve both qualitative and numerical inputs. Existing Gaussian process (GP) methods for handling this mainly assume a different response surface for each combination of levels of the qualitative factors and relate them via a multiresponse cross-covariance matrix. We introduce a substantially different approach that maps each qualitative factor to underlying numerical latent variables (LVs), with the mapped values estimated similarly to the other correlation parameters, and then uses any standard GP covariance function for numerical variables. This provides a parsimonious GP parameterization that treats qualitative factors the same as numerical variables and views them as affecting the response via similar physical mechanisms. This has strong physical justification, as the effects of a qualitative factor in any physics-based simulation model must *always* be due to some underlying numerical variables. Even when the underlying variables are many, sufficient dimension reduction arguments imply that their effects can be represented by a low-dimensional LV. This conjecture is supported by the superior predictive performance observed across a variety of examples. Moreover, the mapped LVs provide substantial insight into the nature and effects of the qualitative factors. Supplementary materials for the article are available online.

ARTICLE HISTORY
Received May 2018
Accepted June 2019

KEYWORDS
Categorical variables; Computer experiments; Metamodeling; Response surface modeling; Surrogate modeling

1. Introduction

Computer simulations play essential roles in today's science and engineering research. As an alternative to more difficult and expensive physical experiments, computer simulations help to explore or experiment with the physical process and understand how input factors affect the response of interest. Gaussian process (GP) models, a.k.a. kriging models, have become the most popular method for modeling simulation response surfaces (Fang et al. 2006; Sacks et al. 1989; Santner et al. 2003). These standard methods for the design and analysis of computer experiments were developed under the premise that all the input variables are quantitative, which fails to describe many applications. For example, consider a stamping operation, in which the response is the maximum strain over a stamped panel, and one of the factors affecting strain is the qualitative lubricant type (e.g., three different types: A, B, and C). Another example is modeling the thermal dynamics of a data center (Qian et al. 2008), which involves qualitative factors such as "hot air return vent location" and "power unit type."

Let $y(\cdot)$ denote the computer simulation response model with inputs $\boldsymbol{w} = (\boldsymbol{x}, \boldsymbol{t}) \in \mathbb{R}^{p+q}$, where $\boldsymbol{x} = \left(x_1, x_2, \ldots, x_p\right)$ represent p quantitative variables, and $\boldsymbol{t} = \left(t_1, t_2, \ldots, t_q\right)$ represent q qualitative factors, with the jth qualitative factor having m_j levels, $j = 1, 2, \ldots, q$. When there are only quantitative inputs $\boldsymbol{x}$, a common model is

$$y(\boldsymbol{x}) = \mu + G(\boldsymbol{x}), \tag{1}$$

where μ is a constant prior mean, $G(\boldsymbol{x})$ is a zero-mean GP with covariance function $K(\cdot,\cdot) = \sigma^2 R(\cdot,\cdot)$, σ^2 is the prior variance, and $R(\cdot,\cdot\,|\boldsymbol{\phi})$ denotes the correlation function with parameters $\boldsymbol{\phi}$. A commonly used correlation function for quantitative variables is the Gaussian correlation function

$$R\left(\boldsymbol{x}, \boldsymbol{x}'\right) = \exp\left\{-\sum_{i=1}^{p} \phi_i \left(x_i - x_i'\right)^2\right\}, \tag{2}$$

which quantifies the correlation between $G(\boldsymbol{x})$ and $G\left(\boldsymbol{x}'\right)$ for any two input locations $\boldsymbol{x} = \left(x_1, \ldots, x_p\right)$ and $\boldsymbol{x}' = \left(x_1', \ldots, x_p'\right)$. $\boldsymbol{\phi} = \left(\phi_1, \ldots, \phi_p\right)$ is the vector of correlation parameters to be estimated via maximum likelihood estimation (MLE), along with μ and σ^2.

These types of correlation functions cannot be directly used with qualitative factors because the distances between the levels of qualitative factors are not defined. To incorporate both qualitative and quantitative factors into GP modeling, a number of covariance structures have been proposed and investigated (McMillan et al. 1999; Joseph and Delaney 2007; Qian et al. 2008; Zhou et al. 2011; Zhang and Notz 2015; Deng et al. 2017). Most methods essentially treat the computer model as a multiresponse GP with a different response for each combination of levels of the qualitative factors, often with some simplifications in the covariance structure. We discuss these methods in detail in Section 3.

In this article, we propose a fundamentally different method of handling qualitative factors in GP models that involves a

CONTACT Daniel W. Apley apley@northwestern.edu Department of Industrial Engineering & Management Sciences, Northwestern University, 633 Clark St, Evanston, IL 60208.

This paper was handled by incoming Editor V. Roshan Joseph throughout the review process.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/TECH.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TECH.

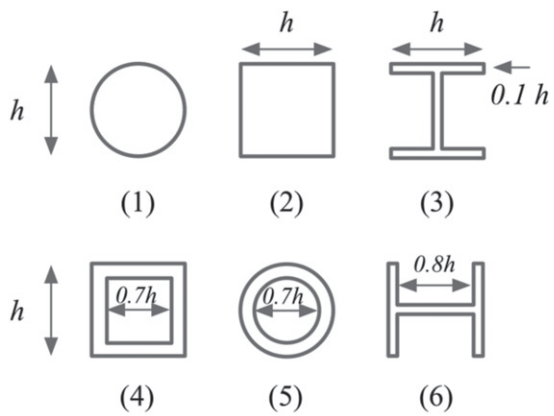

Figure 1. Six cross-sectional shapes, corresponding to six levels of the qualitative factor for the beam bending example: (1) circular cross-section with diameter *h*; (2) square cross-section with height and width *h*; (3) I-shaped cross-section with height and width *h* and thickness 0.1 *h*; (4) hollow square cross-section with outer side length *h* and thickness 0.15 *h*; (5) hollow circular cross-section with outer diameter *h* and thickness 0.15 *h*; (6) H-shape cross-section with height and width *h* and thickness 0.1 *h*.

latent variable (LV) representation of the qualitative factors. The main idea is to map the levels of each qualitative factor to a set of numerical values for some underlying latent unobservable quantitative variable(s). After obtaining this mapping, our GP covariance model over $(\boldsymbol{x}, \boldsymbol{t})$ can be any standard GP covariance model for quantitative variables over $(\boldsymbol{x}, \boldsymbol{z}(\boldsymbol{t}))$, where $\boldsymbol{z}(\boldsymbol{t})$ is the numerical vector of mapped LVs. The mapped values $\{\boldsymbol{z}(\boldsymbol{t})\}$ can be obtained in a straightforward and computationally stable manner via MLE along with the correlation parameters for $\boldsymbol{x}$, and the mapping is scaled so that the correlation parameters for $\boldsymbol{z}$ are unity.

There are strong physical arguments for why our mapped LV approach constitutes a covariance parameterization that, while tractable and involving relatively few parameters to estimate, is flexible enough to capture the behavior of many real physical systems. For any real physical system model having a qualitative input factor, there are *always* underlying physical variables that account for the differences in the response across the different levels of the factor. For example, in the earlier stamping example, differences in the response (panel deformation, strain, stress, etc.) due to different lubricant types *must* be due to the lubricant types having different numerical values (denoted by $\{v_1(t),\ v_2(t),\ v_3(t),\ \ldots\}$) for some underlying physical properties such as lubricity, viscosity, density, and thermal stability of the oil. Otherwise, there is no way to code a simulation model to account for the effects of lubricant type.

To make these arguments more concrete, consider the classic beam bending problem in which the qualitative factor is the cross-sectional shape of the beam with six levels: circular, square, I-shape, hollow square, hollow circular, and H-shape (see Figure 1). The beam has an elastic modulus $E = 600$ GPa and is operating within its linear elastic range. The beam is fixed on one end, and a force of $P = 600$ N is applied vertically at the free end. The response y is the amount of deformation at the free end. In addition to the cross-sectional shape represented by the qualitative factor t, there are two numerical input variables: beam length L and beam width (which is the same as beam height) h. The underlying numerical variables $\{v_1(t), v_2(t), v_3(t), \ldots\}$ for cross-section type t would be the complete cross-sectional geometric positions (normalized by the "size" parameter h of all the elements in the finite element mesh of the beam. The physics of this beam bending problem is transparent enough that we know the beam deflection y depends on the complete high-dimensional geometric descriptors $\{v_1(t), v_2(t), v_3(t), \ldots\}$ via the response function $y(L, h, t) = \frac{L^3}{3\,10^9 h^4 I}$, where $I = I(t) = I(v_1(t), v_2(t), v_3(t), \ldots)$ is the normalized (by h) moment of inertia of the cross-section. Consequently, the underlying high-dimensional variables that govern the effect of the qualitative factor t on y can be mapped down to a single numerical variable $I(t)$.

With advanced knowledge of these physics, one should obviously treat the cross-section shape as the single numerical input $I(t)$. But to illustrate the motivation and justification behind our latent variable Gaussian process (LVGP) approach, suppose such knowledge were unavailable. In this case, one option would be to treat the cross-section as the qualitative factor t and use an existing GP method for qualitative inputs that presumes no underlying numerical structure. A second option would be to include the set of numerical variables $\{v_1(t), v_2(t), v_3(t), \ldots\}$ as inputs, which is not feasible due to the extremely high dimensionality. Our LVGP approach is an attractive alternative that presumes there exists some unknown underlying LV $\boldsymbol{z}(t)$ that captures the joint effect of $\{v_1(t), v_2(t), v_3(t), \ldots\}$ on the response, and the approach attempts to discover the underlying effect of t by estimating the LV mapping $\boldsymbol{z}(t)$. If the approach performs effectively in this case (we demonstrate later that it does), the estimated LV mapping $\boldsymbol{z}(t)$ will represent the normalized moment of inertia $I(t)$.

Returning to the general situation, although there may be many underlying variables $\{v_1(t), v_2(t), v_3(t), \ldots\}$, their collective effect on the response will often be captured by some low-dimensional latent combination of the variables. To see this, notice that their collective effect can always be written as $y = g(\boldsymbol{x}, v_1, v_2, v_3, \ldots)$ for some function $g(\cdot)$. If, for example, the dependence happens to be of the form $y \cong g(\boldsymbol{x}, \beta_1 v_1 + \beta_2 v_2 + \cdots)$, then a single one-dimensional LV $z(t) = \beta_1 v_1(t) + \beta_2 v_2(t) + \cdots$ suffices to capture the effects of the qualitative factor t. More generally, if the dependence happens to be of the form $y \cong g(\boldsymbol{x}, h_1(v_1, v_2, \ldots), h_2(v_1, v_2, \ldots))$ for some functions $g(\cdot)$, $h_1(\cdot)$ and $h_2(\cdot)$, then a two-dimensional LV $\boldsymbol{z}(t) = (h_1(v_1(t), v_2(t), \ldots), h_2(v_1(t), v_2(t), \ldots))$ suffices to capture the effects of the qualitative lubricant type.

The preceding is a rather broad and flexible structure for representing the effects of quantitative variables and qualitative factors on y. For even more general $g(\cdot)$ and more complex dependence of the response on $\{v_1, v_2, \ldots\}$, the same arguments behind sufficient dimension reduction (Cook and Ni 2005; Li 1991) imply that the collective effects of $\{v_1, v_2, \ldots\}$ can be represented approximately as a function of the coordinates over some lower dimensional manifold in the $\{v_1, v_2, \ldots\}$-space. If the manifold is approximately two-dimensional, then $y \cong g(\boldsymbol{x}, h_1(v_1, v_2, \ldots), h_2(v_1, v_2, \ldots))$, and the two-dimensional LV representation that we use in our approach will suffice.

To summarize the justification and advantages of our LVGP approach, it (1) has strong physical justification, since the effect of any qualitative factor t must always be due to a set of underlying numerical variables $\{v_1(t), v_2(t), v_3(t), \ldots\}$, and the effect of these can often be captured by some low-dimensional LV $\boldsymbol{z}(t)$; (2) provides far superior predictive performance, relative to existing alternatives, across the variety of examples that we consider later; (3) has the added benefit of providing excellent interpretability of the effects of a qualitative factor t via inspection of a scatterplot of the two-dimensional (2 D) mapped LV values $\{\boldsymbol{z}(1), \boldsymbol{z}(2), \ldots, \boldsymbol{z}(m)\}$ (m is the number of levels of t) to look for any patterns or clustering in the mapped values; and (4) is flexible in that it allows one's favorite covariance function for numerical variables (e.g., Gaussian, power exponential, Matèrn, lifted Brownian, etc., either separable or nonseparable versions) to be used over the combined original and mapped numerical variables $(\boldsymbol{x}, \boldsymbol{z}(t))$. Advantages (2) and (3) will be demonstrated later in the examples.

The outline of the remainder of the article is as follows. Section 2 describes our LVGP representation of qualitative factors, along with the MLE implementation for estimating all covariance parameters, including the LV mapping $\boldsymbol{z}(\boldsymbol{t})$. Section 3 reviews existing GP models for qualitative and quantitative variables. Section 4 reports numerical comparisons for a number of examples showing that our proposed LVGP method consistently outperforms existing methods and is capable of accurately identifying the underlying LV structure. Section 5 discusses why one would ever treat an input as qualitative if its effect must always be due to underlying numerical variables. Section 6 concludes the article.

2. Latent Variable Representation of Qualitative Factors

2.1. A 1D LVGP Representation for q = 1

We first describe the approach in the context that we have a single qualitative factor t with m levels (labeled $t = 1, 2, \ldots, m$) and are using a one-dimensional (1 D) LV $z(t)$ to represent the m levels. The m levels of t will be mapped to m latent numerical values $(z(1), \ldots, z(m))$ for z. The input $\boldsymbol{w} = (\boldsymbol{x}, t)$ is therefore mapped to $(\boldsymbol{x}, z(t))$, and using the Gaussian correlation function in (2), our correlation model is (a constant prior variance σ^2 is still assumed)

$$\begin{aligned}\mathrm{Cor}\left\{y(\boldsymbol{x}, t), y\left(\boldsymbol{x}', t'\right)\right\} &= \mathrm{Cor}\left\{y(\boldsymbol{x}, z(t)), y\left(\boldsymbol{x}', z\left(t'\right)\right)\right\} \\ &= \exp\left\{-\sum_{i=1}^{p} \phi_i \left(x_i - x_i'\right)^2 - \left(z(t) - z\left(t'\right)\right)^2\right\}, \quad (3)\end{aligned}$$

where ϕ_i's are the correlation parameters for the quantitative variables $\boldsymbol{x}$. Note that there is no correlation parameter for z. We take it to be unity, because when $(z(1), \ldots, z(m))$ are estimated in the MLE optimization, their spacing will appropriately account for the correlation between the levels of the qualitative factor t.

Under the model (3), the log-likelihood function is

$$\begin{aligned}l(\mu, \sigma, \boldsymbol{\phi}, \mathbf{Z}) = &-\frac{n}{2}\ln\left(2\pi\sigma^2\right) - \frac{1}{2}\ln\left|\mathbf{R}(\boldsymbol{\phi}, \mathbf{Z})\right| \\ &- \frac{1}{2\sigma^2}\left(\mathbf{y} - \mu\mathbf{1}\right)^T \mathbf{R}(\boldsymbol{\phi}, \mathbf{Z})^{-1}\left(\mathbf{y} - \mu\mathbf{1}\right), \quad (4)\end{aligned}$$

where n is the sample size, $\mathbf{1}$ is an n-by-1 vector of ones, $\mathbf{y}$ is the n-by-1 vector of observed response values, $\mathbf{Z} = (z(1), \ldots, z(m))$ are the mapped values of the m levels of t, and $\mathbf{R}(\boldsymbol{\phi}, \mathbf{Z})$ is the n-by-n correlation matrix whose elements are obtained by plugging pairs of the n sample values of $(\boldsymbol{x}, t)$ into (3). Without loss of generality, we set the first level $t = 1$ to correspond to the origin in the LV space (i.e., $z(1) = 0$), because in (3) only the relative distances between levels of t in the LV space affect the correlation. For the same reason, fixing $z(1)$ is necessary to prevent indeterminacy or nonidentifiability during the MLE optimization process.

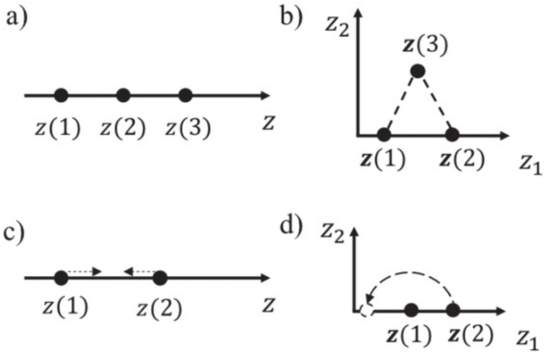

Figure 2. Limitations of 1 D LV representation: (a) in 1 D the LV mappings cannot represent three equally correlated levels; (b) in a 2 D latent space, the three LV mappings $z(1)$, $z(2)$ and $z(3)$ can be arranged as the vertices of an equilateral triangle to represent equal correlations among all three levels; (c) the singularity issue of the covariance matrix when two points become too close to each other when exchanging positions during the MLE optimization search; d) in 2 D, the LVs can move freely to avoid covariance singularity when exchanging positions during the MLE optimization.

In the numerical studies in Section 4, we found that a 1 D latent space effectively captures the correlation structure of qualitative factors in a variety of real and realistic examples. However, using the 1 D latent representation has the following shortcomings, and so we prefer a 2 D latent representation. Suppose the qualitative factor t has three levels and the response correlation $\mathrm{Cor}\left(y(\boldsymbol{x}, t), y\left(\boldsymbol{x}', t'\right) \mid \boldsymbol{\phi}\right)$ for all levels $t \neq t'$ is the same value (e.g., 0.6). To represent this via (3), the three levels must have equal pairwise distances in the LV space, which is impossible using a 1 D representation. This is depicted in Figure 2(a) for the case that $|z(2) - z(1)| = |z(3) - z(2)|$, in which case $|z(3) - z(1)| = 2|z(2) - z(1)|$, so that the correlation between levels $t = 1$ and $t' = 3$ must be smaller than the correlation between the other two pairs of levels. To represent the equal correlation scenario, a 2 D latent space shown in Figure 2(b) is necessary, in which the three 2 D latent mapped values $(\boldsymbol{z}(1), \boldsymbol{z}(2), \boldsymbol{z}(3))$ can form an equilateral triangle. The 2 D latent representation also provides correlation structure flexibility in other regards, beyond what the 1 D representation can provide.

Another potential issue with a 1 D latent representation can occur when the MLE optimizer adjusts the mapped LVs $(z(1), \ldots, z(m))$ along the single latent dimension z. If any two z values become too close at any point in the optimization, this could cause singularity of the correlation matrix. For example, suppose the initial guesses for two latent points (say $z(1)$ and $z(2)$) are reversed from what their MLEs are. As illustrated in Figure 2(c), during the MLE optimization, $z(1)$ and $z(2)$ may need to gradually move toward each other to reverse their positions, which may cause covariance singularity when they

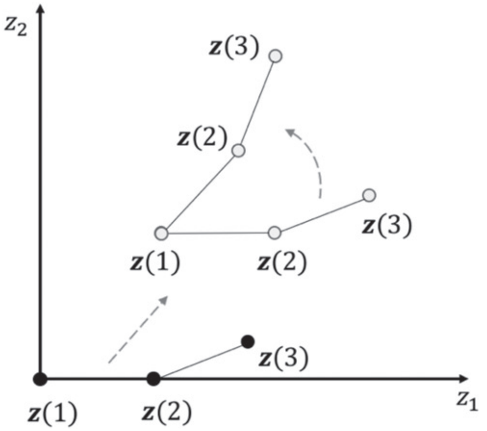

Figure 3. Indeterminacy caused by translation and rotation: Three different configurations for the mapped latent values $\{z(1), z(2), z(3)\}$ have the same pairwise distances and the same covariance structure.

get too close. With more qualitative levels, there is a higher probability of encountering singularity during optimization. In contrast, a 2 D latent space can reduce the likelihood of singularity significantly, because the points can be moved around more freely in the 2 D space. For example, the positions of $z(1)$ and $z(2)$ can be reversed without ever having to move them too close to each other, as shown in Figure 2(d).

2.2. A 2D LV Representation for $q = 1$

As depicted in Figure 2, compared with a 1 D LV representation, a 2 D representation provides more flexibility to capture complex correlation structures for qualitative factors and better numerical behavior in the MLE optimization. With a single qualitative factor ($q = 1$), to extend the model (3) to incorporate a 2 D LV $\boldsymbol{z} = (z_1, z_2) \in \mathbb{R}^2$, we map the m levels of t to the m points $\{\boldsymbol{z}(1) = (z_1(1), z_2(1)), \ldots, \boldsymbol{z}(m) = (z_1(m), z_2(m))\}$ in 2 D latent space. The input $\boldsymbol{w} = (\boldsymbol{x}, t)$ is thus mapped to $(\boldsymbol{x}, \boldsymbol{z}(t))$, and the corresponding Gaussian correlation function is

$$\operatorname{Cor}\{y(\boldsymbol{x}, t), y(\boldsymbol{x}', t')\} = \exp\left\{-\sum_{i=1}^{p} \phi_i (x_i - x_i')^2 - \|\boldsymbol{z}(t) - \boldsymbol{z}(t')\|_2^2\right\}, \quad (5)$$

where $\|\cdot\|_2$ denotes the Euclidean 2-norm. The mapped values for the m levels are again estimated via MLE, along with the other covariance parameters. A total of $2(m-1)-1 = 2m-3$ scalar latent values are required to represent the m different levels of t, because (i) similar to the 1 D case, the first level of t can always be mapped to the origin (i.e., $\boldsymbol{z}(1) = (0, 0)$) to remove the indeterminacy caused by translation invariance; and (ii) to remove indeterminacy due to rotational invariance in the 2 D latent space, we can restrict the 2 D position of the mapped value $\boldsymbol{z}(2)$ for the second level to lie on the horizontal axis. Figure 3 illustrates this for $m = 3$ by showing three different configurations of three mapped latent values $\{\boldsymbol{z}(1), \boldsymbol{z}(2), \boldsymbol{z}(3)\}$ that are translated and rotated versions of each other. They therefore have the same pairwise distances and result in the same covariance structure via (5). Our convention of taking $\boldsymbol{z}(1)$ to be the origin and $\boldsymbol{z}(2)$ to lie on the horizontal axis removes the indeterminacy and reduces the total number of free parameters to estimate from $2m$ to $2m-3$, which scales linearly with the number of levels of the qualitative variable.

With $m > 3$ levels, one might consider a more general version of our approach that uses an $(m-1)$-dimensional LV representation $\boldsymbol{z} = (z_1, \ldots, z_{m-1}) \in \mathbb{R}^{m-1}$ in (5). Similar to the 2 D scenario, to avoid indeterminacy due to rotation/translation invariance, the mapped value for the first level can be taken to be the origin $\boldsymbol{z}(1) = (0, \ldots, 0) \in \mathbb{R}^{m-1}$, and we can likewise restrict $\boldsymbol{z}(2) = (z_1(2), 0, \ldots, 0) \in \mathrm{R}^{m-1}$, $\boldsymbol{z}(3) = (z_1(3), z_2(3), 0, \ldots, 0) \in \mathbb{R}^{m-1}$, …, and $z(m) = (z_1(m), z_2(m), \ldots, z_{m-1}(m)) \in \mathbb{R}^{m-1}$. This model would require estimating $m(m-1)/2$ independent mapped LV parameters in total, which is the same as in the unrestrictive covariance model of Qian et al. (2008). This $m-1$ dimensional LVGP model is a very general covariance structure that allows the independent representation of all $m(m-1)/2$ pairwise correlations of the response across the m qualitative levels for t. However, we do not believe such a general $(m-1)$-dimensional LV representation is needed for most problems. This is supported by the conceptual “sufficient dimension reduction” arguments given in Section 1 and the numerical results in Section 4, for which the correlation structures of the qualitative factors have effective low-dimensional (1 D or 2 D) representations.

2.3. A 2D LV Representation with Multiple Qualitative Factors

In general, suppose there are $q > 1$ qualitative factors $\boldsymbol{t} = (t_1, t_2, \ldots, t_q)$, where the jth factor $t_j \in \{1, 2, \ldots, m_j\}$, and m_j denotes the number of levels of t_j. Our approach has a very efficient and natural way to handle multiple qualitative factors, that is, akin to how multiple numerical input variables are handled in GP modeling. We simply use a different 2 D LV $\boldsymbol{z}^j$ to represent each qualitative factor t_j ($j = 1, 2, \ldots, q$). As explained earlier, there are $2m_j - 3$ parameters for each $\boldsymbol{z}^j$, so that the total number of parameters is only $\sum_{j=1}^{q}(2m_j - 3)$.

The corresponding Gaussian correlation function for our approach is

$$\operatorname{Cor}\left\{y(\boldsymbol{x}, \boldsymbol{t} = (t_1, \ldots, t_q)), y\left(\boldsymbol{x}', \boldsymbol{t}' = \left(t_1', \ldots, t_q'\right)\right)\right\} = \exp\left\{-\sum_{j=1}^{p} \phi_j \left(x_j - x_j'\right)^2 - \sum_{j=1}^{q} \left\|\boldsymbol{z}^j(t_j) - \boldsymbol{z}^j\left(t_j'\right)\right\|_2^2\right\}, \quad (6)$$

where $\boldsymbol{z}^j(l) = (z_1^j(l), z_2^j(l))$ denotes the 2 D mapped LV for level l of the qualitative factor t_j. The $2m_j - 3$ values for the mapped LVs for each factor t_j, along with the parameters $\boldsymbol{\phi}$, μ, and σ^2 of the GP model, are estimated via MLE.

In addition to yielding a relatively parsimonious yet flexible parameterization, this approach also has the following desirable characteristic. Most of the existing GP approaches for qualitative factors treat the computer model as a multiresponse GP with a different response for each combination of levels. In contrast, our LVGP model treats the response at different level combinations to be from a single response surface that is continuous over the numerical LVs that account for the effects of the qualitative factors. This, together with using a separate 2 D LV $\boldsymbol{z}^j$ to represent each qualitative factor, results in an approach that is consistent with how numerical input variables are handled in standard GP modeling. Moreover, even though we have used the separable (in $\boldsymbol{x}$ and $\boldsymbol{z}$) Gaussian covariance in (6), any covariance model used for numerical variables can be used over the joint $(\boldsymbol{x}, \boldsymbol{z})$ space. This includes either the separable or nonseparable versions of the power exponential, Matèrn (Rasmussen et al. 2006), and lifted Brownian (Plumlee and Apley 2017) covariance functions.

3. Review of Existing GP Approaches for Qualitative Factors

3.1. *Unrestrictive Covariance (UC)*

A popular approach in the literature for GP modeling with qualitative variables was introduced by Qian et al. (2008) and further developed in Zhou et al. (2011). They assumed

$$\operatorname{Cor}\left\{y(\boldsymbol{x}, t), y\left(\boldsymbol{x}', t'\right)\right\} = \tau_{t,t'} \exp\left\{-\sum_{i=1}^{p} \phi_i \left(x_i - x_i'\right)^2\right\}, \quad (7)$$

where $\tau_{t,t'}$ is the correlation between the responses corresponding to level t and t'. An $m \times m$ correlation matrix $\boldsymbol{\tau}$ with row-t, column-t' entry $\tau_{t,t'}$ is used to represent the correlations across all m levels of the qualitative variable. To ensure that the correlation defined in (7) is valid, the matrix $\boldsymbol{\tau}$ must be positive definite with unit diagonal elements (PDUDE). When there are $q > 1$ qualitative factors, one approach is to define a single qualitative factor that represents combinations of levels of all the qualitative factors and then use (7). Alternatively, a somewhat less general structure that was also considered in Qian et al. (2008) is the Kronecker product structure

$$\begin{aligned}&\operatorname{Cor}\left\{y\left(\boldsymbol{x}, \boldsymbol{t} = \left(t_1, \ldots, t_q\right)\right), y\left(\boldsymbol{x}', \boldsymbol{t}' = \left(t_1', \ldots, t_q'\right)\right)\right\}\\ &\quad = \prod_{j=1}^{q} \tau^{j}_{t_j, t_j'} \exp\left\{-\sum_{i=1}^{p} \phi_i \left(x_i - x_i'\right)^2\right\}, \end{aligned}\quad (8)$$

where $\tau^{j}_{l,l'}$ represents the correlation between levels l and l' of t_j. Zhou et al. (2011) later simplified the estimation procedure for $\boldsymbol{\tau}$ to ensure positive definiteness by using a hypersphere decomposition (Rebonato and Jäckel 1999).

Zhang and Notz (2015) showed that one could use indicator variables in the Gaussian correlation function to generate the correlation structure in (8). For positive integers i, l, and l', define the level indicator functions

$$I_l(i) = \begin{cases} 1 & i = l \\ 0 & i \neq l \end{cases} \quad (9)$$

and

$$W_{l,l'}(i) = \begin{cases} I_l(i) + I_{l'}(i) & \text{if } l \neq l' \\ I_l(i) & \text{if } l = l' \end{cases}, \quad (10)$$

and consider the correlation function

$$\begin{aligned}&\operatorname{Cor}\left\{y\left(\boldsymbol{x}, t = \left(t_1, \ldots, t_q\right)\right), y\left(\boldsymbol{x}', t' = \left(t_1', \ldots, t_q'\right)\right)\right\}\\ &\quad = \prod_{j=1}^{q} \exp\left\{-\sum_{l,l'=1}^{m_j - 1} \phi^{j}_{l,l'} \left(W_{l,l'}\left(t_j\right) - W_{l,l'}\left(t_j'\right)\right)^2\right\}\\ &\quad \times \exp\left\{-\sum_{i=1}^{p} \phi_i \left(x_i - x_i'\right)^2\right\}, \end{aligned}\quad (11)$$

where $\left\{\phi^{j}_{l,l'} : 1 \leq l, l' \leq m_j - 1\right\}$ are additional parameters to be estimated via MLE. Zhang and Notz (2015) showed that (11) is equivalent to (8) for $\tau^{j}_{l,l'} > 0$, in that there is a one-to-one correspondence between the $\phi^{j}_{l,l'}$'s in (11) and the $\tau^{j}_{l,l'}$'s in (8). Using the formulation in (11) allows one to use standard GP fitting packages to estimate the $\tau^{j}_{l,l'}$'s in the Qian et al. (2008) method with a mild restriction that $\tau^{j}_{l,l'} > 0$. When a single qualitative factor ($q = 1$) is used to represent all the combinations of levels of multiple qualitative factors, (11) reduces to (7) with all $\tau_{l,l'} > 0$. There is no restriction on the elements of $\boldsymbol{\tau}$ in (7) and (8) as long as it is a PDUDE, so it is sometimes referred to as the unrestrictive covariance (UC). Because of symmetry, there are $m(m-1)/2$ free parameters to be estimated in $\boldsymbol{\tau}$, which represent all $m(m-1)/2$ pairwise correlations of the qualitative factor levels.

3.2. *Multiplicative Covariance*

Qian et al. (2008) also discussed some simplified special cases of the UC model. The simplest model assumes $\tau_{l,l'} = \tau$ for all $l \neq l'$, which is referred to as an exchangeable covariance (EC) (Joseph and Delaney 2007; Qian et al. 2008). Another simplified model termed the multiplicative covariance (MC) (McMillan et al. 1999; Qian et al. 2008) assumes that for all $t \neq t'$

$$\tau_{t,t'} = e^{-\left(\theta_t + \theta_{t'}\right)}, \quad (12)$$

where θ_l is a parameter associated with level l of the qualitative factor t, and there are m parameters in this model. As pointed out in Zhang and Notz (2015), this method is equivalent to using a standard GP for quantitative variables with the qualitative variable represented by the set of indicator variables in (9), analogous to how nominal categorical variables are handled in linear regression.

When $m \leq 3$ the MC model is nearly equivalent to the UC model, with the only difference being that $\tau_{t,t'}$ are restricted to being nonnegative (Zhang and Notz 2015). However, when $m \geq 4$, the MC model has the following undesirable properties, as shown in Zhang and Notz (2015). Suppose $m = 4$ and the response surfaces (over $\boldsymbol{x}$) for levels 1 and 2 are highly correlated, the response surfaces for levels 3 and 4 are highly correlated, but the response surfaces for levels 1 and 2 are very different from the surfaces for levels 3 and 4. According to (12), since each $\theta_l > 0$, in order to make $\tau_{1,2} \approx \tau_{3,4} \approx 1$, we must have θ_1, θ_2, θ_3, and θ_4 all close to 0. But in this case, the correlation between levels 1 and 3 becomes $\tau_{1,3} = e^{-(\theta_1 + \theta_3)} \approx 1$, which contradicts the assumption that levels 1 and 3 are not correlated. The MC model fails in this case because it uses only m parameters to specify

$m\,(m-1)\,/2$ pairwise correlations, and the simplified parameterization fails to capture this common physical situation. In contrast, our LVGP model can easily handle this case even with the 1 D LV representations via setting $z\,(2) \approx 0$ and $z\,(3) \approx z\,(4) \gg 0$. We believe that our simplified parameterization using LVs is more consistent with many physical systems and generally is more effective at capturing commonly occurring correlation structures, while still requiring only a small number of parameters.

3.3. *Additive GP Model with Qualitative Variables*

The UC and MC models both assume multiplicative forms of correlations across the quantitative factors and qualitative factors. Deng et al. (2017) proposed the additive covariance structure

$$\text{Cov}\left\{y\,(\boldsymbol{x},\boldsymbol{t})\,,y\left(\boldsymbol{x}',\boldsymbol{t}'\right)\right\} = \sum_{j=1}^{q} \sigma_j^2 \tau^j_{t_j,t'_j} R\left(\boldsymbol{x},\boldsymbol{x}' \,|\, \boldsymbol{\phi}^{(j)}\right), \tag{13}$$

where $R(\boldsymbol{x},\boldsymbol{x}'|\boldsymbol{\phi}^{(j)})$ is the Gaussian correlation function defined in (2) with correlation parameters $\boldsymbol{\phi}^{(j)}$ associated with the qualitative factor t_j, σ_j^2 is a prior variance term associated with qualitative factor t_j, and $\tau^j_{t_j,t'_j}$ has the same definition as in (8). This covariance model is equivalent to assuming $y\left(\boldsymbol{x},\ t_1,\ldots,t_q\right) = \mu + G_1\,(\boldsymbol{x},\ t_1) + \cdots + G_q\left(\boldsymbol{x},\ t_q\right)$,where μ is the overall mean, and the G_j's are independent zero-mean GPs, each with covariance functions over $\left(\boldsymbol{x},\ t_j\right)$ given by the individual terms in (13). When there is only one qualitative factor, this model is equivalent to the covariance model (8). For $q > 1$, Deng et al. (2017) argued that it provides more flexibility for modeling complex computer simulations than the model (8), which assumes a fixed covariance structure over $\boldsymbol{x}$ for all qualitative factors.

It should be noted that if all categorical inputs have two levels, our LVGP covariance (6), the Qian et al. (2008) Kronecker product covariance (8), and the MC covariance (12) are all equivalent to the standard GP approach for numerical inputs but using binary numerical coding for the two-level categorical inputs. Hence, we focus on the situation of more than two levels for the categorical inputs.

4. Numerical Comparisons

In this section, we conduct numerical studies to investigate the effectiveness of the proposed LVGP model (6) with a Gaussian correlation function on a number of examples. The supplementary materials section provides additional examples and further details on the examples in this section. We compare the proposed method with the three covariance structures reviewed in the previous section:

(a) UC covariance in (8) (Qian et al. 2008; Zhou et al. 2011), using the equivalent reformulation (11) discussed in Zhang and Notz (2015);

(b) MC covariance (12) (McMillan et al. 1999; Qian et al. 2008; Zhang and Notz 2015), using the equivalent reformulation (9) with indicator variables discussed in Zhang and Notz (2015);

(c) Additive GP with unrestrictive correlation (Add_UC) defined in (13), which is equivalent to UC when there is only a single qualitative factor.

The Gaussian correlation function in (2) is used for all quantitative variables $\boldsymbol{x}$ in these four methods. To evaluate the model accuracy of each method, we use the relative root-mean-squared error (RRMSE) for the fitted GP model predictions over $N = 10{,}000$ hold-out test points:

$$\text{RRMSE} = \sqrt{\frac{\sum_{i=1}^{N}\left(\hat{y}\,(\boldsymbol{w}_i) - y\,(\boldsymbol{w}_i)\right)^2}{\sum_{i=1}^{N}\left(y\,(\boldsymbol{w}_i) - \bar{y}\right)^2}}, \tag{14}$$

where $\hat{y}(\boldsymbol{w}_i)$ and $y\,(\boldsymbol{w}_i)$ denote the predicted and the true values, respectively, at the input test location $\mathbf{w}_i$, and $\bar{y}$ is the average of the true responses over the 10,000 test points. The 10,000 test points are generated uniformly for both the quantitative variables and qualitative factors. For each example, we used 30 replicates, where on each replicate we generated a different "training" design, the data from which were used to fit the four covariance models, and then we calculated the resulting test RRMSE for the four models. Each training design was a maximin Latin hypercube design (LHD) in the quantitative variables, with the levels of the qualitative factors randomly sampled. The design sizes were chosen so that the RRMSE for the best model for each example was less than 0.1, to ensure that the designs were of sufficient size to allow reasonable prediction accuracy.

We fit the UC and MC models through the same optimization routine for MLE used in our LVGP model. MATLAB code from the supplemental materials of Deng et al. (2017) was used to fit the Add_UC model. To have a common basis for comparison, when fitting all models, we used 200 random initial guesses for the GP hyper-parameters to help ensure good MLE solutions (one might wish to use more initial guesses for higher dimensional problems). During optimization, the correlation parameters for all quantitative inputs are reparametrized as $\theta_i = log_{10}(\phi_i)$, with $\theta_i \in [-3, 3]$, and each LV $z_i^j\,(l)$ is restricted to the interval $[-2, 2]$. Formerly, we had used a much larger interval $[-10, 10]$ over which to search for the MLEs of the LVs. However, their MLEs were almost always much smaller than this, so we now restrict the search range to $[-2, 2]$. This typically allows sufficiently small correlations between levels, when small correlations are needed. An LHD is used for generating the 200 random initial guesses to cover the search space as evenly as possible. For the MLE optimization, we use the MATLAB function *fmincon*, which uses an interior-point method with BFGS for a Hessian approximation.

4.1. *Beam Bending Example Revisited*

When applying our LVGP approach to the beam bending example discussed in the introduction, we do not incorporate the physics knowledge that the underlying numerical variables $\{v_1(t),\ v_2\,(t)\,,\ v_3(t),\ \ldots\}$ characterizing the cross-section impact the response only via the normalized moment of inertia $I(t)$. Instead, we rely on the LVGP approach to discover the underlying LV structure. As discussed in Section 1, if our LVGP approach performs effectively, the estimated LV mapping $\boldsymbol{z}(t)$

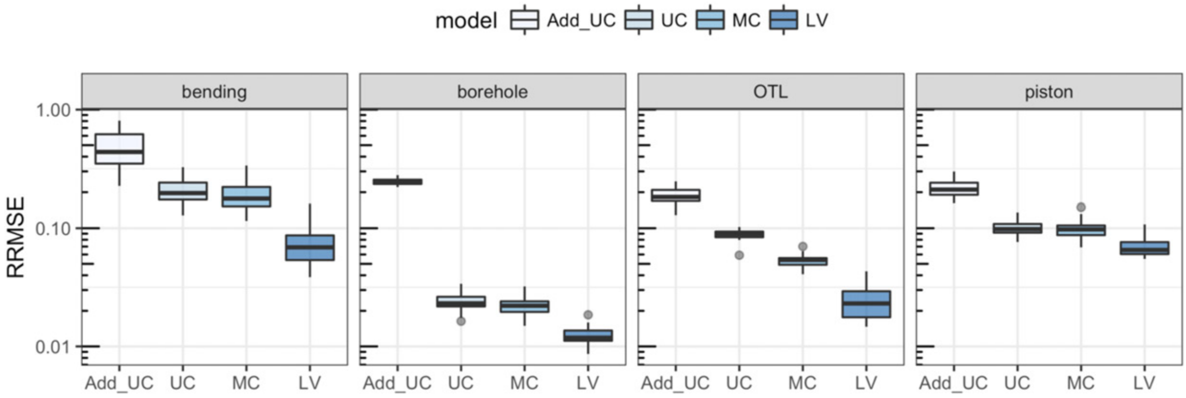

Figure 4. Boxplots of RRMSE across 30 replicates for the four engineering examples with $n = 60$, 80, 60, and 100, respectively. Our LVGP model achieves the smallest RRMSE in each example. Note that the y-axis is in log scale.

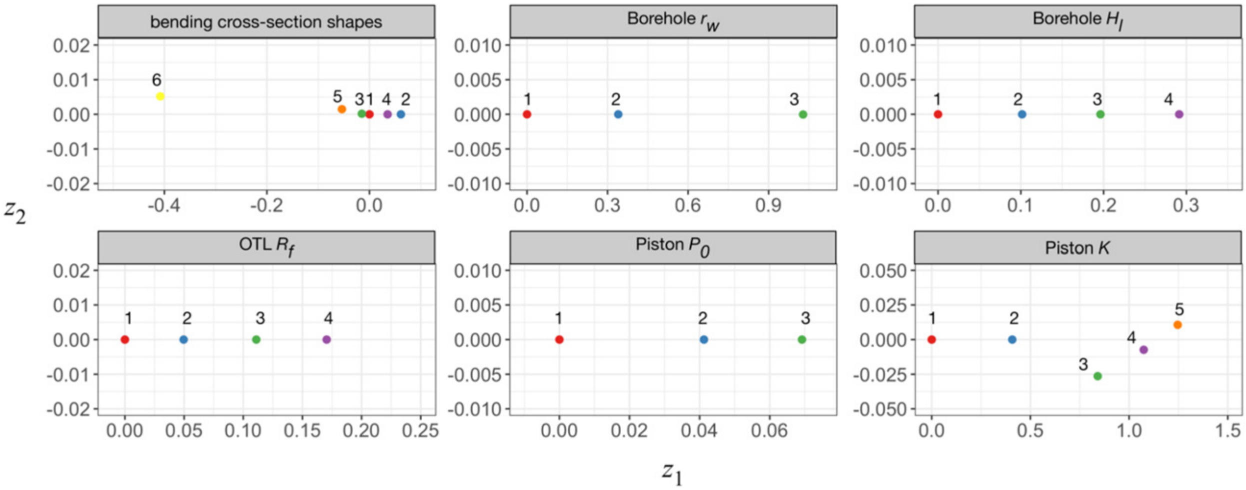

Figure 5. Estimated 2 D LVs $z = (z_1, z_2)$ representing the levels of the qualitative factors in the four engineering examples for a typical replicate: the values of z_2 are small compared with z_1, indicating that the estimated latent representation is a one-dimensional representation that closely matches the settings in Table S2 of the supplementary materials.

will represent the normalized moment of inertia $I(t)$. From basic mechanics, the normalized moments of inertia for the six cross-sections in Figure 1 are $I_1 = \pi/64 = 0.0491$, $I_2 = 1/12 = 0.0833$, $I_3 = 0.0449$, $I_4 = 0.0633$, $I_5 = 0.0373$, and $I_6 = 0.0167$, and their inverses (which turn out to be very closely related to the mapping $\boldsymbol{z}(t)$) are $1/I_1 = 20.4$, $1/I_2 = 12.0$, $1/I_3 = 22.3$, $1/I_4 = 15.8$, $1/I_5 = 26.8$, and $1/I_6 = 59.9$. Notice that the H-shaped cross section (level $t = 6$) has substantially different $1/I$ than the other cross-sections, and the six cross sections ordered from largest to smallest $1/I$ are levels 6, 5, 3, 1, 4, then 2. The ordering and relative spacing agrees nearly perfectly with the estimated LVs $\boldsymbol{z}\,(1)$—$\boldsymbol{z}\,(6)$ shown in Figure 5. Consequently, our LVGP model correctly discovered the underlying mapped LV $\boldsymbol{z}(t)$ that captures the effect of the qualitative factor on y.

Figure 4 shows boxplots of the RRMSE over the 10,000 hold-out test prediction points across 30 replicates. On each replicate, a different maximin LHD was generated, and each of the four models was refit. Our LVGP model had substantially better RRMSE performance than the other covariance models in this beam bending example and also across the other three examples of real engineering models (borehole, OTL, and piston models) described in the supplementary materials. The MC and UC models performed similarly to each other, except that MC worked a little better than UC on the OTL example. The Add_UC model had the highest error across all four examples, perhaps because these real engineering examples do not have the additive structure that it assumes.

In addition, Figure 5 shows that the estimated LVs are positioned nearly exactly along the horizontal z_1 axis for all four examples. Because there truly was a single latent numerical variable associated with each qualitative factor t_j in all of these examples, and the ordering and relative distances between the numerical values of the mapped qualitative levels in Figure 5 closely mimic those for the true levels (see Table S2 in the supplementary materials), the LVGP approach has effectively identified the underlying latent numerical structure for each example.

4.2. A Materials Design Example with Qualitative Inputs

As emphasized throughout this article, all qualitative factors in physics-based simulations must impact the response via some underlying numerical variables $\{v_1(t),\ v_2(t),\ v_3(t),\ \ldots\}$. However, in many situations, the underlying numerical variables may be so high-dimensional and the simulation physics so complex that it precludes conveniently identifying them and incorporating them into a GP model with only numerical variables. This is the case in the following materials design example (Balachandran et al. 2016). The dataset consists of the simulated shear modulus (the response, y) of material compounds belonging to the family of M_2AX phases. The M atom has ten levels (i.e., 10 different candidate choices for the compound) {Sc, Ti, V, Cr, Zr, Nb, Mo, Hf, Ta, W}, the A atom has two levels {C, N}, and the X atom has twelve levels {Al, Si, P, S, Ga, Ge, As, Cd, In, Sn, Tl, Pb}. Thus, there are three qualitative factors with 10, 2, and 12 levels, respectively, to represent the different choices of atoms for the compound. Among the total 240 possible combinations, 17 combinations have negative shear modulus and thus are not considered in this example (see Balachandran et al. 2016 for more details).

In the original study, the authors considered GP surrogate modeling. However, due to the high dimensionality and lack of transparency of the underlying $\{v_1(t),\ v_2(t),\ v_3(t),\ \ldots\}$, and due to the lack of effective GP modeling software for qualitative inputs, the authors used a GP model for numerical-only inputs with a relatively small set of numerical features (which can be viewed as a small subset of $\{v_1\ (t)\ ,\ \mathrm{v}_2(t),\ v_3(t),\ \ldots\}$) that they suspected would have large effects on the response. In total, they chose seven features to serve as their numerical GP inputs, which are the *s*-, *p*-, and *d*-orbital radii for the M atom, and the *s*- and *p*-orbital radii for the A and X atoms. The orbital radii are from the Waber-Cromer scale. We refer to their GP modeling approach with only these seven numerical inputs as the "Quant_only" approach.

In the following, we show the advantages of using GP modeling with the original three qualitative inputs over the Quant_only GP model, and we also show the advantages of the LVGP model over existing GP models that can handle qualitative factors. We consider two versions of the LVGP model: One using only the three qualitative inputs (denoted as LV_qual), and the other using the three qualitative inputs in addition to the seven orbital radii numerical variables (denoted as LV). The seven numerical variables are in some sense redundant if the three qualitative inputs are included since the latter are functions of the former. However, one might speculate that there may be advantages to include them along with the qualitative inputs if they truly have a large impact on the response. The other three models that we compare are three existing GP models that we discussed in Section 3 to handle qualitative and quantitative inputs (ADD_UC, UC, and MC), all with the three qualitative inputs plus the seven numerical inputs.

There are 223 data points in total, and we used 200 of them for training and the remaining 23 to compute the test RRMSE. The training and test sets were chosen randomly from the 223 points, and we repeated this procedure for ten replicates, where on each replicate we chose different random subsets to serve as the training and test sets and repeated the modeling. Figure 6 shows that our LVGP methods (both LV and LV_qual) have much lower RRMSE than any of the other approaches. Notice that Quant_only results in consistently large RRMSE, possibly due to the seven chosen numerical features providing an insufficient quantitative representation of the effects of the qualitative levels. Although it includes the qualitative factors along with the quantitative features, the UC approach does not improve the accuracy compared with Quant_only. This is likely due to the fact that two of the qualitative variables have relatively large numbers of levels (10 and 12), resulting in a large number of parameters to estimate in the UC model. Both MC and ADD_UC have better RRMSE than UC, although our LVGP model achieves even better RRMSE. The best performing model was LV_qual since its 25th, 50th, and 75th RRMSE percentiles were all slightly better than those for the LV model, and substantially better than all other models. It is somewhat surprising that the LV_qual model performed better than the LV model, since the additional seven numerical features included in the LV model were speculated to have large effects. The benefit of including the additional seven numerical features appears to be offset by the additional challenge of estimating more hyperparameters. We view this as evidence that our LVGP approach handles qualitative factors and identifies the underlying LV structure effectively.

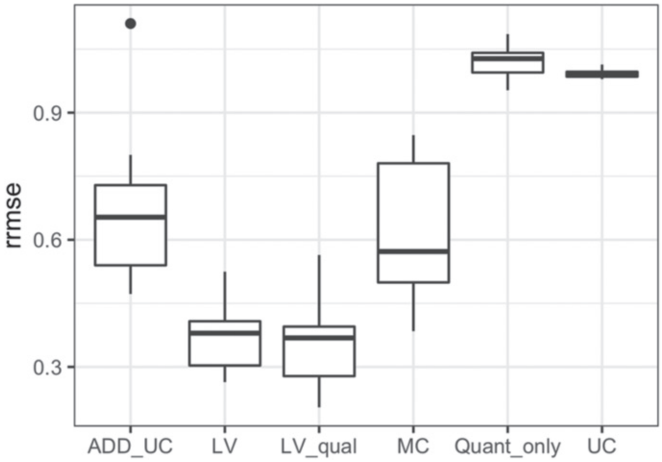

Figure 6. RRMSE comparison for the materials design example. Our LVGP method with (LV) or without (LV_qual) the additional seven numerical features achieved the lowest RRMSE.

This example illustrates an important reason why one would consider using qualitative factors in a GP model, even though their effects must be due to underlying numerical variables: Without definitive prior knowledge and a simulator whose mechanisms are transparent, selecting an appropriate set of low-dimensional features of the high-dimensional $\{v_1(t),\ v_2(t),\ v_3(t),\ \ldots\}$ is often subjective and provides an incomplete representation of the effects of the qualitative t (as witnessed from the poor Quant_only performance in Figure 6). If we instead work with the qualitative factors as inputs, the LVGP model can account for the more complete information not captured by quantitative variable features, thereby improving the GP model predictions.

4.3. Borehole Example with a True Latent Space That is 2D

In the beam bending example, the effects of the qualitative factor can be reduced to a function of a single underlying numerical variable, so that the true latent numerical space for the qualitative factor is 1 D. Here, we modify the borehole example described in the supplementary materials by creating

Table 1. Mapping from the 2 D underlying numerical variables (r_w, H_l) to the single qualitative factor t in the revised borehole example.

Level of t	r_w	H_l	Level of t	r_w	H_l	Level of t	r_w	H_l
1	0.05	700	5	0.10	700	9	0.15	700
2	0.05	740	6	0.10	740	10	0.15	740
3	0.05	780	7	0.10	780	11	0.15	780
4	0.05	820	8	0.10	820	12	0.15	820

a qualitative factor t having 12 levels that represent 12 discrete combinations of two underlying numerical variables r_w and H_l (see the supplementary materials). The mapping from (r_w, H_l) to the level of t is listed in Table 1. The other quantitative input variables all have the same ranges shown in Table S1 of the supplementary materials. This example represents the case where multiple underlying numerical variables vary across the levels of a qualitative factor, and we demonstrate below that our LVGP model can successfully reveal the underlying structure, just as it did in the Figure 5 examples.

Figure 7(a) plots the estimated 2 D LVs associated with the qualitative factor t, from which we see that the 12 levels of t are arranged into three groups, each representing a different level of r_w. Moreover, within each group, as H_l increases, the points move predominantly along the z_1 direction. Thus, the estimated 2 D LVs have successfully revealed the dependence of the qualitative factor on the underlying numerical variables r_w and H_l, with z_2 approximately representing r_w, and z_1 approximately representing a combination of H_l and r_w.

Notice that the levels of r_w and H_l are evenly spaced in their original units, as shown in Table 1, but the estimated z_1 and z_2 values are not evenly spaced in Figure 7(a). The reason is that in our LVGP model, the distances between LVs depend on the response correlation across the qualitative levels, which depends not only on the distances between the underlying inputs but also on the behavior of the response. The response surface contour plot in Figure 7(b) further illustrates the reason: when r_w is at its lower level 0.05, the response does not change as much along the H_l dimension as when r_w is at its higher level 0.15. Consequently levels 1–4 are more closely spaced in Figure 7(a) than levels 9–12 are. In this sense, our LVGP model has correctly identified the structural dependence of the qualitative levels on a set of underlying numerical variables, in terms of capturing the response similarities/differences across the levels of the factor.

4.4. LVGP vs. BNGP, and the Effect of Dimensionality

This example compares our LVGP approach (in which only the qualitative levels of the input are available) with an approach that treats the underlying numerical variables $\{v_1(t),\ v_2\ (t)\ ,\ v_3(t),\ \ldots\}$ as available and uses them in a standard GP model for numerical inputs. We refer to the latter as the benchmark numerical GP (BNGP) approach, since it uses information (the underlying numerical variables) that is not used in the LVGP approach and that might not be easily available in practice. The same design of experiments is used for both methods, so that the numerical variables used in the BNGP approach are only evaluated at locations corresponding to the qualitative levels.

For this example, we replace the qualitative variable t with $m = 5$ in Math Function 1 described in (S1) of the supplementary materials by a set of J underlying numerical variables $\{v_1(t),\ v_2\ (t)\ ,\ v_3(t),\ \ldots,v_J(t)\}$ and investigate the effect of dimensionality J on the performance. Specifically, the response is

$$y\left(\boldsymbol{x}, v_1(t), v_2(t), \ldots v_J(t)\right) = 7\sin\left(2\pi x_1 - \pi\right) + \left[J^{-1/2}\sum_{j=1}^{J} v_j(t)\right] \times \sin\left(2\pi x_2 - \pi\right). \tag{15}$$

To be consistent with the $y(\boldsymbol{x}, t)$ response surface in (S1) of the supplementary materials, we chose the $5 \times J$ values for $\{v_1(t),\ v_2(t),\ \ldots\ v_J(t)\colon\ t = 1, 2, \ldots, 5\}$ so that $\{J^{-1/2}\sum_{j=1}^{J} v_j(t)\colon\ t = 1, 2, \ldots, 5\} = \{1,\ 13,\ 1.5,\ 9.0,\ 4.5\}$. Beyond that, we randomly generated the values for $\{v_1(t),\ v_2\ (t)\ ,\ \ldots\ v_J(t)\colon\ t = 1, 2, \ldots, 5\}$. More specifically, we used the basis vectors $\mathbf{A} = \left[\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_J\right]$ for the J-dimensional space, where $\mathbf{a}_1 = J^{-1/2}\mathbf{1}$, and $\mathbf{a}_j = J^{-1/2}\left(J-1\right)^{-1/2}\left(J\mathbf{e}_j - \mathbf{1}\right)$ for $j = 2,\ 3,\ \ldots,\ J$, with $\mathbf{1}$ and $\mathbf{e}_j$ denoting the J-length column vector of ones and the J-length column vector of zeros with a one in the jth position, respectively. Then, for each $t = 1, 2, \ldots, 5$, we used $[v_1(t),\ v_2(t),\ \ldots\ v_J(t)]^T = \mathbf{A}[v(t),\ u_2(t),\ \ldots\ u_J(t)]^T$ with $\{v(t)\ :\ t = 1, 2, \ldots, 5\} = \{1,\ 13,\ 1.5,\ 9.0,\ 4.5\}$ and the $5 \times (J-1)$ values for $\{u_2(t),\ u_3(t),\ \ldots\ u_J(t)\ :\ t = 1, 2, \ldots, 5\}$ randomly generated from a uniform distribution over the interval $[0, 10]$.

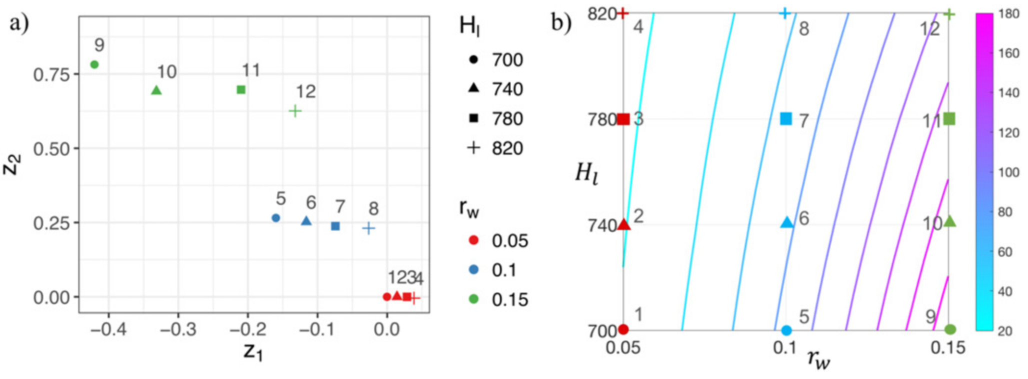

Figure 7. (a) estimated 2 D mapped LVs representing the 12 levels of the qualitative factor t in the revised borehole example. The latent representation successfully uncovered the structural dependence of the factor levels on the two underlying numerical variables: the three levels of r_w (represented by colors) are distributed along z_2 dimension and within each r_w group the four levels of H_l correspond to z_1 varying; (b) contour plot of the response in revised borehole example as a function of r_w and H_l with the other numerical variables in Table S1 fixed at their mean values.

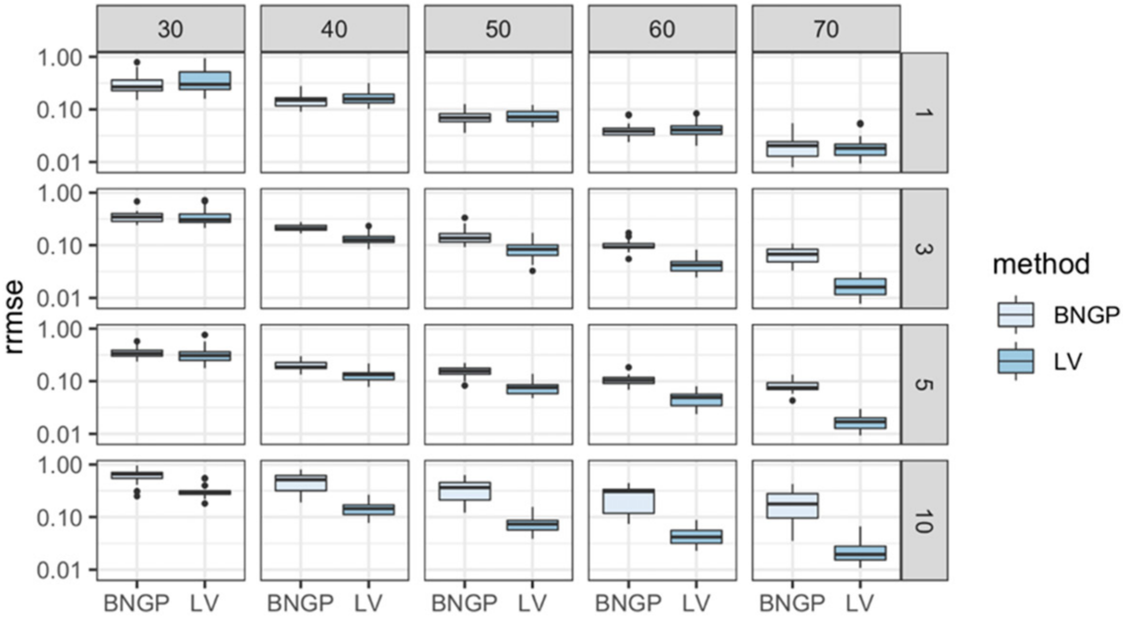

Figure 8. RRMSE comparison (each boxplot is for 20 replicates) of BNGP vs. LVGP for the example in (15) with design sizes $n = 30, 40, 50, 60$, and 70 (corresponding to columns) when the dimension of the underlying numerical variables is $J = 1, 3, 5$, and 10 (corresponding to rows). Our LVGP model has only slightly higher error than the benchmark BNGP model that uses the underlying numerical $v(t)$ when $J = 1$, and the differences decrease with larger n. The BNGP model degrades significantly when the dimension of the underlying numerical variables increases.

We conducted 20 replicates of the example, where on each replicate we generated a different set of $5 \times (J - 1)$ uniform random numbers for $\{u_2(t), u_3(t), \ldots u_J(t) : t = 1, 2, \ldots, 5\}$ and a different design of experiments. For the latter, we generated a size-n LHD in the $\{x_1, x_2\}$ space and then assigned the level for t for each of the n runs by randomly sampling one of its five levels. The BNGP model was fit to the same data as the LVGP model but using the underlying numerical $\{v_1(t), v_2(t), \ldots v_J(t)\}$ instead of t. Figure 8 compares the RRMSEs across 20 replicates for five different DOE sizes (n = 30, 40, 50, 60, and 70) and for four different values of J (1, 3, 5, 10).

The main conclusion drawn from Figure 8 is that if the underlying numerical variables are low-dimensional ($J = 1$), very little accuracy is lost if we use the LVGP approach, relative to using the BNGP approach that incorporates the numerical variable information; and if the underlying numerical variables are higher dimensional ($J \geq 3$), the LVGP approach gives much better accuracy than the BNGP approach. We note that for $J = 1$ and the smaller designs ($n < 50$ roughly), the BNGP approach does indeed perform slightly better than the LVGP approach, but the difference becomes negligible for the larger designs ($n >$ 50 roughly).

5. Why Use Qualitative Factors at All?

Aside from the superior numerical performance of the LVGP approach demonstrated in our examples, the main justification for LVGP is the recognition that the effects of qualitative factors on a numerical response must always be due to some set of underlying numerical variables $\{v_1(t), v_2(t), v_3(t), \ldots\}$ that vary across the different levels of the factor. In light of this, one may question whether it would be better to simply identify what are the variables $\{v_1(t), v_2(t), v_3(t), \ldots\}$ that vary across the levels t of the factor, and then to include these as numerical inputs in a standard GP model for only numerical variables. Identifying the numerical variables should generally be straightforward, albeit perhaps tedious, since whoever coded the simulation must know which variables he/she included in the code. Assuming this can be done in practice, the appropriateness of a purely numerical GP model largely depends on the dimension of $\{v_1(t), v_2(t), v_3(t), \ldots\}$ and on the level of prior knowledge regarding how they collectively affect the response variable.

The adverse effect of dimensionality of $\{v_1(t), v_2(t), v_3(t), \ldots\}$ on predictive performance when they are treated as purely numerical was demonstrated in Figure 8. To further elucidate the issue, reconsider the beam bending example, in which $\{v_1(t), v_2(t), v_3(t), \ldots\}$ for the cross-sectional qualitative factor are the complete set of 2 D coordinates for every integration point in the finite element mesh of the cross-section. If 1,000 integration points are used, then there are 2,000 underlying numerical variables $\{v_1(t), v_2(t), v_3(t), \ldots, v_{2000}(t)\}$ that vary as the level t (cross-section shape) varies. The best way to handle this is to have prior knowledge of the physics of the system and to know in advance that $\{v_1(t), v_2(t), v_3(t), \ldots, v_{2000}(t)\}$ only affect the response via the 1 D moment of inertia variable $I(t) = I\left(v_1(t), v_2(t), v_3(t), \ldots, v_{2000}(t)\right)$. In this case, it would be naive to treat the cross-sectional shape as a qualitative factor. Instead, one would represent the cross-sectional shape via the single numerical variable $I(t)$ and include it in a standard GP model for numerical-only inputs (although the $J = 1$ results in Figure 8 indicate that one might not lose too much if the LVGP approach is used).

Such strong prior knowledge of how $\{v_1(t), v_2(t), v_3(t), \ldots\}$ affect the response is not generally available. If $\{v_1(t), v_2(t), v_3(t), \ldots\}$ is low-dimensional (e.g., only one or two variables), then one should probably forego a qualitative factor treatment and, instead, include $\{v_1, v_2, v_3, \ldots\}$ as additional numerical variables in the simulation experiments. This would entail varying $\{v_1, v_2, v_3, \ldots\}$ over some experiment designed for numerical variables, conducting the simulation runs at these values, and then using a GP model for numerical variables to model the response surface.

On the other hand, if $\{v_1(t), v_2(t), v_3(t), \ldots\}$ is high-dimensional, it may be impossible to include them all as additional numerical variables in the simulation. This is clearly the case for the beam bending example, for which one would never attempt to include $\{v_1, v_2, v_3, \ldots, v_{2000}\}$ as 2000 additional numerical variables in the simulation experiment and in the GP surrogate model. Instead, with only six different levels, one would be far better off treating the cross section as a qualitative factor and using the LVGP approach. An additional benefit of the LVGP approach is that it can help to discover the low-dimensional LVs $\{z_1(t), z_2(t)\} = \{z_1(v_1(t), v_2(t), v_3(t), \ldots), z_2(v_1(t), v_2(t), v_3(t), \ldots)\}$ that capture the effects of the underlying high-dimensional variables $\{v_1(t), v_2(t), v_3(t), \ldots\}$ on the response. This was clearly evident from Figures 5 and 7.

In general, it may be better to forego a qualitative factor GP model and instead represent them as numerical inputs in a standard GP model if either (*i*) there are only a few underlying numerical variables that differ across the levels of the qualitative factor or (*ii*) there are many underlying numerical variables, but one has strong prior knowledge that they collectively affect the response only via a few low-dimensional combinations, and the functional forms of these combinations are known. If many underlying numerical variables differ across levels, and one does not understand the physics clearly enough to identify a few low-dimensional combinations on which the response depends, then a GP model with qualitative inputs should be used.

6. Conclusions

In this article, we developed an LVGP model for GP-based simulation response surface modeling with both quantitative and qualitative factors. The approach maps the qualitative factor levels to a corresponding set of 2 D latent numerical variable values so that distances in the LV space account for response correlations across levels of the qualitative factors. We argue that the proposed two-dimensional LVGP model (6) is flexible enough to accurately capture complex correlations of many qualitative factors. To support this, we have (1) demonstrated consistently superior predictive performance across a variety of mathematical and engineering examples (Figures 4, 6, and S1) and (2) provided a physical explanation of why differences in the response behavior across qualitative factor levels are truly due to underlying numerical variables that can be mapped down to a lower dimensional space of LVs (e.g., the beam bending example).

Another desirable characteristic of our LVGP approach is that the estimated LVs provide insight into the relationship between the levels of a factor, regarding how similar or different the response surfaces are for the different levels. In all of our examples, the visualization of the LV space (Figures 5, 7, and S2) successfully revealed the structure of the true underlying variables that account for the response differences between levels. Moreover, in contrast to the existing methods for handling qualitative factors that were reviewed in Section 3, our LVGP approach is compatible with any standard GP correlation function, including nonseparable correlation functions such as power exponential, Matèrn and lifted Brownian. This allows greater flexibility when modeling complex systems. The resulting covariance function in our LVGP model always results in a valid (positive semidefinite) covariance matrix without having to incorporate additional constraints, making the MLE routine easier to implement.

Our numerical performance studies focused on the RRMSE comparisons and on the LVGP model's ability to estimate the underlying LV structure. However, one often uses GP models for their built-in ability to quantify the uncertainty in the response predictions via prediction intervals obtained from the built-in mean squared prediction error formula or some appropriate computational Bayesian analysis. Further studies on whether the LVGP model has relative advantages or disadvantages for uncertainty quantification would be useful.

Finally, our focus has been on GP modeling for simulation response surfaces. In the much broader landscape of general regression modeling with qualitative factors, it still holds that the effects of any qualitative factor t on a *numerical* response variable must be due to some underlying set of characteristics that differ across the levels of t. If these characteristics are all quantifiable, then we can view them as the underlying numerical variables $\{v_1(t), v_2(t), v_3(t), \ldots\}$ to which we have referred throughout this article. If one accepts this viewpoint, then the LV mapping concepts in this article may have much broader applicability in regression with qualitative variables than just the GP modeling setting considered in this article. As a simple example, the standard linear regression approach of encoding a qualitative factor with m levels as $m - 1$ 0/1 dummy variables can be viewed as mapping the qualitative factor down to a 1 D LV numerical space, where the mapped LV values are exactly the estimated regression coefficients of the 0/1 dummy variables. However, the LV interpretations are not as clear when one considers interactions between the 0/1 dummy variables and other numerical or qualitative predictors. We are currently investigating whether some of the LV mapping concepts that we have used in this article may also be useful in the broader regression context.

Acknowledgments

The authors thank the Editor (Roshan Joseph) who handled this article and the anonymous Associate Editor and two Referees for their many helpful suggestions that have improved the article and its presentation.

Funding

This work was supported in part by National Science Foundation (grant nos. CMMI-1537641, DMREF grant no. 1729743), and the NIST-ChiMaD (Center for Hierarchical Materials Design) Grant.

Supplementary materials

The online supplementary materials for this article contain numerical performance results for several additional examples, as well as further details on the examples in Section 4. The R-package “LVGP”, which is available from the Comprehensive R Archive Network (CRAN) at *http://CRAN.R-project.org/package=LVGP*, contains the code for fitting LVGP models to general mixed-variable datasets.

References

Balachandran, P. V., Xue, D., Theiler, J., Hogden, J., and Lookman, T. (2016), “Adaptive Strategies for Materials Design Using Uncertainties,” *Scientific Reports*, 6, 19660. [298]

Cook, R. D., and Ni, L. (2005), “Sufficient Dimension Reduction Via Inverse Regression: A Minimum Discrepancy Approach,” *Journal of the American Statistical Association*, 86, 316–327. [292]

Deng, X., Lin, C. D., Liu, K. W., and Rowe, R. K. (2017), “Additive Gaussian Process for Computer Models With Qualitative and Quantitative Factors,” *Technometrics*, 59, 283–292. [291,296]

Fang, K., Li, R. Z., and Sudjianto, A. (2006), *Design and Modeling for Computer Experiments*, Computer Science and Data Analysis Series. [291]

Joseph, V. R., and Delaney, J. D. (2007), “Functionally Induced Priors for the Analysis of Experiments,” *Technometrics*, 49, 1–11. [291,295]

Li, K. C. (1991), “Sliced Inverse Regression for Dimension Reduction,” *Journal of the American Statistical Association*, 86, 316–327. [292]

McMillan, N. J., Sacks, J., Welch, W. J., and Gao, F. (1999), “Analysis of Protein Activity Data by Gaussian Stochastic Process Models,” *Journal of Biopharmaceutical Statistics*, 9, 145–160. [291,295,296]

Plumlee, M., and Apley, D. W. (2017), “Lifted Brownian Kriging Models,” *Technometrics*, 59, 165–177. [295]

Qian, P. Z. G., Wu, H., and Wu, C. F. J. (2008), “Gaussian Process Models for Computer Experiments With Qualitative and Quantitative Factors,” *Technometrics*, 50, 383–396. [291,294,295,296]

Rasmussen, C. E., Williams, C. K. I., Processes, G., Press, M. I. T., and Jordan, M. I. (2006), *Gaussian Processes for Machine Learning*. Cambridge, MA: MIT Press. [295]

Rebonato, R., and Jäckel, P. (1999), “The Most General Methodology to Create a Valid Correlation Matrix for Risk Management and Option Pricing Purposes,” *QUARC Preprint*, 1–12. [295]

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989), “Design and Analysis of Computer Experiments,” *Statistical Science*, 4, 409–423. [291]

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer, p. 286. [291]

Zhang, Y., and Notz, W. I. (2015), “Computer Experiments With Qualitative and Quantitative Variables: A Review and Reexamination,” *Quality Engineering*, 27, 2–13. [291,295,296]

Zhou, Q., Qian, P. Z. G., and Zhou, S. (2011), “A Simple Approach to Emulation for Computer Models With Qualitative and Quantitative Factors,” *Technometrics*, 53, 266–273. [291,295,296]