SUPPLEMENTARY MATERIALS FOR

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

These supplementary materials contain a number of additional examples comparing the predictive performance of the LVGP model with other methods and investigating its ability to discover the correct LV mapping structure. They also contain further details on some of the examples considered in Section 4. Here, we first test the four covariance models considered in Section 4 on two mathematical functions that have been used in the literature as benchmark problems involving qualitative factors (Deng et al. 2017; Swiler et al. 2014). We also include descriptions of the three additional engineering examples (in addition to the beam bending example) for which performance results are summarized in Figures 4 and 5 of the paper, which are popular examples from the literature for assessing surrogate models with numerical inputs. In these examples, we have converted some of the numerical input variables to qualitative factors. This has the benefit of providing a second means of assessing the effectiveness of the LVGP covariance model. Namely, since the qualitative factors in this case are truly due to some underlying numerical variables, and since we know what values of the numerical variable correspond to the factor levels, we can compare the true values with our estimated LV values. Unless otherwise noted, we use the same simulation setup as described in Section 4 of the paper.

Math Function 1

The first mathematical test function is adapted from Swiler et al. (2014) and has one qualitative variable t with five levels, and two continuous variables $x_1, x_2 \in [0,1]$.

This function has regions where the response behaviors at different qualitative levels are very similar. The different levels of t are associated with the coefficients of the second term in the following definition of the function;

$$y(x, t) = \begin{cases} 7\sin(2\pi x_1 - \pi) + \sin(2\pi x_2 - \pi) & if\ t = 1 \\ 7\sin(2\pi x_1 - \pi) + 13\sin(2\pi x_2 - \pi) & if\ t = 2 \\ 7\sin(2\pi x_1 - \pi) + 1.5\sin(2\pi x_2 - \pi) & if\ t = 3 \\ 7\sin(2\pi x_1 - \pi) + 9.0\sin(2\pi x_2 - \pi) & if\ t = 4 \\ 7\sin(2\pi x_1 - \pi) + 4.5\sin(2\pi x_2 - \pi) & if\ t = 5 \end{cases} \quad (S1)$$

The levels of t therefore have a true ordering 1-3-5-4-2, which is obtained by comparing the coefficients of the second terms in (S1). To fit the four GP models, a different training set of size 70 is generated for each of the 30 replicates, using maximin LHDs for the quantitative variables x_1 and x_2 with the levels of t randomly assigned. The left panel of Figure S1 shows the prediction accuracy over the 10,000 hold-out points via boxplots of the 30 RRMSE values across the 30 replicates. The median values of the RRMSE for the ADD_UC, UC, MC, and LVGP models are 0.181, 0.103, 0.134, and 0.015, respectively. Thus, our LVGP model achieved an average RRMSE that was roughly an order of magnitude smaller than for the other models in this example.

In addition to more accurate response predictions, our LVGP model can provide valuable insight into the effects of the qualitative factor on the response. For example, the left panel of Figure S2 displays the estimated 2D LVs for a typical replicate for this example. Even though the 2D LVs were estimated, their MLE values fell almost exactly on a straight line. The straight line corresponds to the z_1 axis, since $z(1)$ is restricted to the origin, and $z(2)$ is restricted to falling on the z_1 axis. This result is desirable,

since the qualitative variable truly corresponds to a 1D LV in this example, as discussed above. Moreover, the estimated LVs are correctly ordered as 1-3-5-4-2 with levels 1 and 3 positioned very close to each other. This is very consistent with (S1), in which the response surfaces at level 1 and 3 have the smallest differences. The response surface at level 2 has the most substantial differences with the surface at level 1, and our estimated $z(2)$ is correctly positioned as the farthest from the origin. It should be noted that the estimated LVs do have non-zero values in the z_2 coordinate, although they are so small that they are visually indiscernible.

Math Function 2

The second function used for comparing the four models is from Deng et al. (2017) with $p = 5$ quantitative variables and $q = 5$ qualitative factors, each having three levels:

$$y = \sum_{i=1}^{5} \frac{x_i(t_{6-i}-2)}{80} + \prod_{i=1}^{5} \cos\left(\frac{x_i}{\sqrt{i}}\right) \sin\left(\frac{50(t_{6-i}-2)}{\sqrt{i}}\right), \tag{S2}$$

where $-100 \leq x_i \leq 100$, for $i = 1, \ldots, p$, and t_j ($j = 1, \ldots, 5$) are the five qualitative factors, each having three levels $\{1, 2, 3\}$. We generated a maximin LHD of size $n = 100$ as the training set for the x_i's in each of the 30 replicates, with the levels of the t_j's randomly assigned. The qualitative factors affect the response y in a nearly additive manner, and as a result, the additive GP model outperforms the UC and MC models, as seen in the right panel of Figure S1. However, our proposed model works even better than the additive GP model, having the lowest median RRMSE value of 0.045, which is roughly four times smaller than the additive GP model's RRMSE of 0.185. The estimated LVs associated with t_1 are shown in Figure S2 for a typical

replicate, for which the three levels are approximately equally spaced along the z_1 axis and correctly ordered as 1-2-3, which again agrees very closely with the true underlying numerical t_j in (S2). The estimated LVs for t_2 to t_6 were in similar agreement and are not shown in Figure S2.
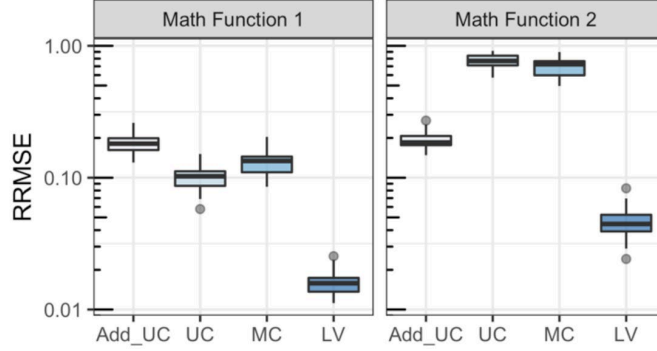

Figure S1: Boxplots of RRMSE over 30 replicates for the two mathematical functions with design sizes of $n = 70$ and 100, respectively. Our LVGP approach outperforms the other three methods. Note that the y-axis is in log scale.
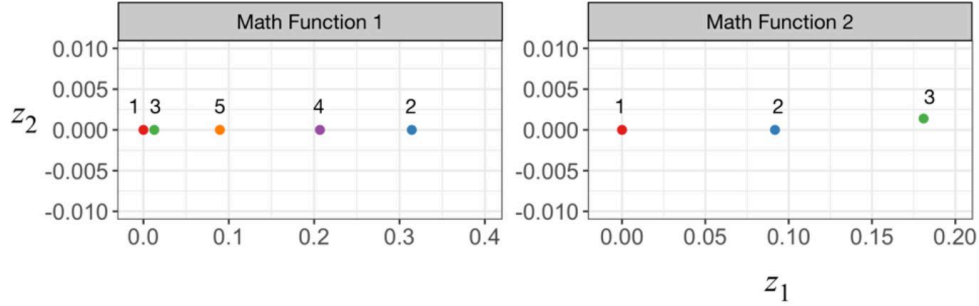

Figure S2: Estimated 2D LVs $\mathbf{z} = (z_1, z_2)$ representing the levels of the qualitative factors in the two mathematical test functions for typical replicates: the values of z_2 are nearly zero for both examples, indicating the true 1D latent structure was correctly identified; and the estimated spacing of between the latent values match with the settings in each example.

Borehole Example

A commonly-used function to study computer simulation surrogate modeling is the

borehole function, $y = 2\pi T_u (H_u - H_l) \left\{ \log\left(\frac{r}{r_w}\right)\left(1 + 2\frac{LT_u}{\log\left(\frac{r}{r_w}\right)r_w^2 K_w} + \frac{T_u}{T_l}\right)\right\}^{-1}$, where

the 8 inputs are $(T_u, r, r_w, H_u, T_l, H_l, L, K_w)$. See Morris et al. (1993) for a full

description of the variables. We treat r_w and H_l as qualitative factors with three and

four levels, respectively. The qualitative factor levels for this and the subsequent

examples are listed in Table S2 .

OTL Example

The midpoint voltage of a transformerless (OTL) circuit function is $y =$

$B\frac{(V_{b1}+0.74)(R_{c2}+9)}{B(R_{c2}+9)+R_f} + 11.35\frac{R_f}{B(R_{c2}+9)+R_f} + 0.74B\frac{R_f}{R_{c1}}\frac{R_{c2}+9}{B(R_{c2}+9)+R_f}$, where $V_{b1} = 12R_{b2}/$

$(R_{b1} + R_{b2})$, and the inputs are $(R_{b1}, R_{b2}, R_f, R_{c1}, R_{c2}, B)$. See Ben-Ari and Steinberg

(2007) for details. We treat R_f and B as qualitative factors having 4 and 6 levels,

respectively.

Piston Example

This example models the cycle time for a piston moving within a cylinder as $y =$

$2\pi\sqrt{\frac{M}{k+S^2\frac{P_0V_0T_a}{V^2 T_0}}}$, where $V = \frac{S}{2k}\sqrt{A^2 + 4k\frac{P_0}{T_0}T}$ and $A = P_0S + 19.62M - \frac{kV_0}{S}$. The

inputs are $(M, S, V_0, k, P_0, T_a, T_0)$. See Sacks et al. (1989) for details. We treat the two

variables P_0 and k as qualitative factors each having 3 and 5 levels, respectively.

Table S1: Quantitative input ranges for the four engineering examples

Bending	Borehole	OTL circuit	Piston
$L \in [10, 20]$	$r \in [100, 50000]$	$R_{b1} \in [50, 150]$	$M \in [30, 60]$
$h \in [1, 2]$	$T_u \in [63070, 115600]$	$R_{b2} \in [25, 70]$	$S \in [0.005, 0.020]$
	$H_u \in [990, 1110]$	$R_{cf} \in [1.2, 2.5]$	$V_0 \in [0.002, 0.010]$
	$T_l \in [63.1, 116]$	$R_{c2} \in [0.25, 1.20]$	$T_a \in [290, 296]$
	$L \in [1120, 1680]$		$T_0 \in [340, 360]$
	$K_w \in [9855, 12045]$		

Table S2: Qualitative factors and their levels for the four engineering examples

Level	Bending t	Borehole $t_1 = r_w$	Borehole $t_2 = H_l$	OTL circuit $t_1 = R_f$	OTL circuit $t_2 = B$	Piston $t_1 = P_0$	Piston $t_2 = k$
1	Circular	0.05	700	0.5	50	9000	1000
2	Square	0.10	740	1.2	100	10000	2000
3	I-shape	0.15	780	2.1	150	11000	3000
4	Hollow Circular		820	2.9	200		4000
5	Hollow Square				250		5000
6	H-shape				300		

Table S1 lists the ranges of all quantitative input variables in the four engineering examples (beam bending, borehole, OTL circuit, Piston), and Table S2 displays the qualitative factors and their levels. The performance results are summarized in Figures 4 and 5 of the paper.

Example with a 10-D $\{v_1(t), v_2(t), v_3(t), ...\}$ that Cannot be Reduced to a 2-D $z(t)$

Throughout, we have used a two-dimensional LV $z(t)$ for each qualitative factor, based in part on sufficient dimension reduction arguments that a two-dimensional $z(t)$ should provide a reasonable approximation of the effects of the underlying

$\{v_1(t), v_2(t), v_3(t), ...\}$ for many qualitative factors. This was illustrated with the beam bending example, in which $\{v_1(t), v_2(t), v_3(t), ...,v_{2000}(t)\}$ is very high-dimensional, but their collective effect on the response is captured via the single one-dimensional combination $I(t) = I\big(v_1(t), v_2(t), v_3(t), ..., v_{2000}(t)\big)$.

The following example is a modification of Math Function 2 in which there are ten underlying numerical variables $\{v_1(t), v_2(t), v_3(t), ...,v_{10}(t)\}$ associated with a single qualitative factor t, but their effects cannot be captured exactly by a two-dimensional $\mathbf{z}(t)$. The response function is

$$y\big(\mathbf{x}, v_1(t), v_2(t), ... \, v_{10}(t)\big) = \sum_{i=1}^{10} \frac{x_i v_{11-i}}{4000} + \prod_{i=1}^{10} \cos\left(\frac{x_i}{\sqrt{i}}\right) \sin\left(\frac{v_{11-i}}{\sqrt{i}}\right), \qquad \text{(S3)}$$

where $-100 \le x_i \le 100$, and $-50 \le v_i \le 50, i = 1, ...,10$. The qualitative factor t has 5 levels, and we used the following mapping between t and $\{v_1(t), v_2(t), ... \, v_{10}(t)\}$. We randomly generated all 50 values for $\{v_1(t), v_2(t), ... \, v_{10}(t): t = 1,2, ...,5\}$ uniformly within $[-50, 50]$. We used 20 replicates, and on each replicate, a different set of 50 values for $\{v_1(t), v_2(t), ... \, v_{10}(t): t = 1,2, ...,5\}$ were generated. For the design of experiment, on each replicate we generated a different size-n LHD in the $\{x_1, x_2, ... \, x_{10}\}$ 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 $\{v_1(t), v_2(t), ... \, v_{10}(t)\}$ instead of t. As discussed in Section 4 of the paper, the BNGP method uses knowledge (of $\{v_1(t), v_2(t), ... \, v_{10}(t)\}$) that is not used by the LVGP method. Figure S3 shows the RRMSE comparison of the two models with different numbers of starting points for hyperparameter estimation (24 and

120) and training design sizes ($n = 80$ and 100). Even though the LVGP model uses a two-dimensional $\mathbf{z}(t)$, it achieved similar errors as the benchmark BNGP model and was evidently a reasonable approximation. The fact that the LVGP model performed as well as the BNGP model in this special situation that we contrived to be unfavorable to the LVGP model is strong evidence of its effectiveness.
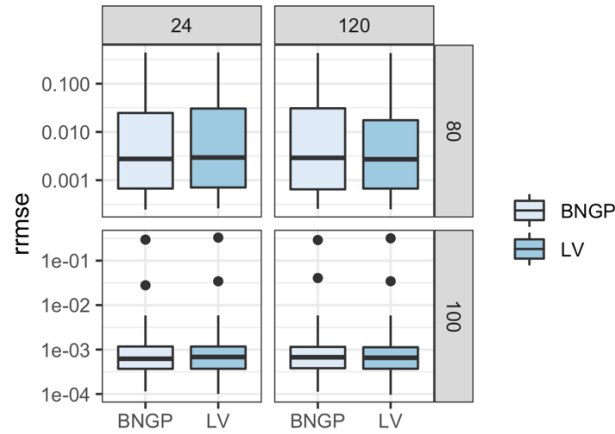

Figure S3: RRMSE comparison for the example in (S3) across 20 replicates. The two columns represent 24 and 120 starting points for hyperparameter estimation, while the two rows represent design sizes $n = 80$ and 100. Even though this example was contrived to be unfavorable to the LVGP model, in all cases it achieved similar errors as the BNGP approach that uses the underlying numerical $\{v_1(t), v_2(t), \dots v_{10}(t)\}$.