APPENDIX A
SYNTHETIC DATASETS

The synthetic datasets are generated based on the real-world MIMIC-III dataset. We move two consecutive times of a time series closer, if the relative difference $\Delta\tilde{x}$ in two consecutive measurements is smaller than the relative difference $\Delta\tilde{t}$ in two consecutive times. The relative differences $\Delta\tilde{x}$ and $\Delta\tilde{t}$ of a time series are given by

$$\Delta\tilde{x}_i = \frac{|x_i - x_{i-1}|}{\sum_{i=2}^{B} |x_i - x_{i-1}|}$$
$$\Delta\tilde{t}_i = \frac{|t_i - t_{i-1}|}{\sum_{i=2}^{B} |t_i - t_{i-1}|}.$$

The scaling factor $d \in (0, 1)$ controls how much we move times. If $d = 0$, we do not move times. In other words, the synthetic dataset at $d = 0$ is the same as the real-world MIMIC-III dataset. As d increases, stronger constraints are introduced to synthetic data. The synthetic time t' for a time series is generated as follows:

$$t'_i = \begin{cases} t_1, & \text{if } i = 1 \\ t_i + \sum_{j=2}^{i} [d(\Delta\tilde{x}_j - \Delta\tilde{t}_j)S], & \text{otherwise} \end{cases}$$
$$S = \sum_{j=2}^{B} (|t_j - t_{j-1}|).$$

If a time series has missing values, we first calculate the synthetic times for the observed measurements. Then we perform a linear interpolation between real times and synthetic times for observed measurements to generate synthetic times for missing measurements.

APPENDIX B
RESULTS

A. Performance Comparison

Table III shows a variable-wise comparison of the imputation models on the real-world MIMIC (d=0) dataset. Our two imputation models outperform all comparison models on all variables. MixMI is better than MixMI-LL on most variables. All models except GP and MTGP achieve a much lower error on Hematocrit and Hemoglobin than on other variables. The reason is that these two variables are highly correlated. Those methods that capture the correlation between variables can reasonably infer missing values for Hematocrit from observed measurements of Hemoglobin, and vice versa. Compared to MICE, MixMI achieves even lower errors on these two variables, which indicates that temporal correlations captured by our model help to make better estimation of missing values, even when there is a more dominant cross-sectional correlation.

As shown in Table II, all models except MTGP benefit from the increment of d, the scaling factor used to generate synthetic data. The reason is that all models take into account temporal aspects and the measurements in the synthetic time series have stronger temporal correlations as d increases. MICE and GMM also benefit from the temporal correlations because we include time as a feature in addition to variable features. We also try to exclude times, however, experiments show that these two models perform better when times are included. MixMI-LL is outperformed by 3D-MICE when d increases to 1, while MixMI shows its robustness to the variation of d in our current experimental settings.

To evaluate the interpretability of our model, we conduct an experiment where our imputation model is trained and evaluated on the real-world MIMIC dataset with one variable being removed at a time. We calculate the performance decrease, i.e. the increase of MASE, of our model with a given variable being removed against the model trained with all variables. The more increase the more important a variable is. The comparison of importance is shown in Table IV.

We compare the running times of all models on the real-world MIMIC dataset. All models run on the same Linux server, in parallel if possible, using up to 20 2.20GHz cores. MixMI-LL (taking 4.2 hours), GP (1.1 hours), MTGP (1.2 hours) GMM (2.2 hours) and M-RNN (0.9 hours) are the fastest models; 3D-MICE (156.1 hours) is the slowest; MICE (77.5 hours) and MixMI (109.5 hours) are in the middle.

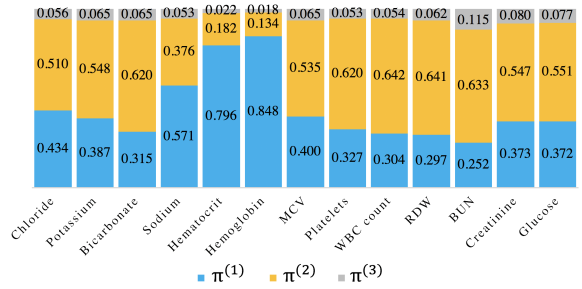


Fig. 3. Component weights comparison on real-world MIMIC dataset

B. Component Weights

The key information our imputation model uses to estimate missing values is the component weight π, which quantifies the interaction of cross-sectional and temporal correlations. We study MixMI on the real-world MIMIC dataset and visualize how the interaction is captured by our model. Fig. 3 shows a comparison of all component weights for variables across all patients and times. When imputing Hematocrit and Hemoglobin, our model relies mostly (79.6% and 84.8%) on the linear model through the cross-sectional view because of the strong correlation between these two variables. Interestingly, most values of $\pi^{(3)}$ are lower than 10%, which indicates that temporal correlations are not strong in these variables in our dataset. However, our model is still able to detect the week temporal correlations and utilizes them to improve imputation. Besides, that $\pi^{(2)}$ has a larger value in most of the variables attests the importance of the linear component through temporal view, removing which in our experiments can cause a 10% performance decrease. Additionally, we observe that when predicting the missing values at the beginning and the end of a time series, our model reasonably uses a lower value of $\pi^{(3)}$ than in the middle. On average, $\pi^{(3)}$ of the first and

TABLE III
MASE ON THE REAL-WORLD MIMIC DATASET BY VARIABLE AND IMPUTATION MODEL. THE BOLD NUMBERS ARE THE BEST VALUES AMONG ALL IMPUTATION MODELS.

Variable	GP	MTGP	M-RNN	GMM	MICE	3D-MICE	MixMI-LL	MixMI
Chloride	0.12993	0.12549	0.10865	0.10650	0.10575	0.10836	0.08664	**0.08603**
Potassium	0.11533	0.11256	0.11222	0.10963	0.10997	0.10822	0.09453	**0.09442**
Bicarbonate	0.13196	0.12905	0.12371	0.12231	0.12275	0.11984	0.10302	**0.10254**
Sodium	0.12525	0.11939	0.11557	0.10159	0.10138	0.10727	**0.08787**	0.08807
Hematocrit	0.11436	0.10780	0.09189	0.06518	0.06558	0.06726	0.05486	**0.05482**
Hemoglobin	0.14168	0.12997	0.06556	0.05774	0.05772	0.06301	0.05117	**0.05103**
MCV	0.14215	0.13732	0.13798	0.13391	0.13474	0.13340	**0.11634**	0.11657
Platelets	0.13855	0.13087	0.14369	0.14203	0.14236	0.12815	0.10090	**0.10070**
WBC count	0.13963	0.13614	0.14583	0.14026	0.14068	0.13060	0.10934	**0.10913**
RDW	0.14592	0.13668	0.15938	0.15778	0.15836	0.13897	0.11340	**0.11340**
Blood urea nitrogen (BUN)	0.12358	0.12720	0.16345	0.15160	0.15189	0.11814	0.09479	**0.09410**
Creatinine	0.13341	0.12803	0.14904	0.13211	0.13212	0.12217	0.10067	**0.10014**
Glucose	0.12491	0.12420	0.11677	0.11771	0.11794	0.11921	**0.10493**	0.10501

TABLE IV
IMPORTANCE OF VARIABLES TO OUR IMPUTATION RESULTS ON THE REAL-WORLD MIMIC DATASET, MEASURED IN NORMALIZED INCREASE

Variable	MASE increase (%)	Normalized increase
Hematocrit	4.52	1
Hemoglobin	3.47	0.76
Chloride	2.11	0.46
Sodium	2.05	0.45
BUN	0.82	0.17
Creatinine	0.79	0.16
Bicarbonate	0.61	0.12
MCV	0.44	0.08
Potassium	0.42	0.08
WBC count	0.34	0.06
Platelet	0.33	0.06
RDW	0.27	0.05
Glucose	0.06	0

the last time indices is 13.9% lower than those in the middle. This is due to the usage of GPs, which usually produce less confident estimates at the end points of series. Furthermore, we observe an increment of $\pi^{(3)}$ as we impose stronger temporal correlations in synthetic datasets, which further validates the ability of our model in capturing such interaction.

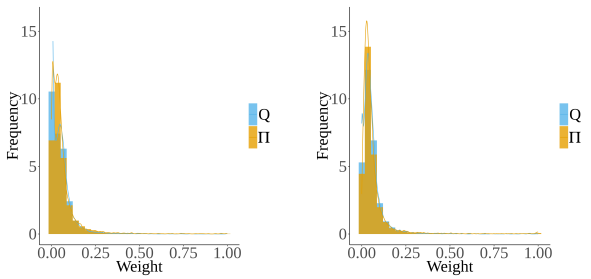
C. Individualized Weights

By introducing individualized (per patient) mixing weights Π defined in (10), we improve the performance in MASE score from 0.08351 to 0.07538, an improvement of 9.73% compared against the model where each mixture component has a fixed weight for all patient cases. The reason that individualized weights are better than fixed weights in our model might be that they better approximate the responsibilities Q defined in (12).

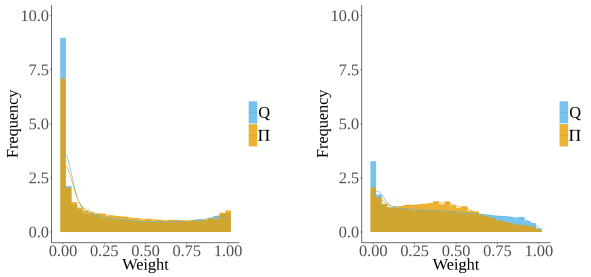
In training, we can optimize the responsibility a component should take to "explain" an observed target value $x_{p,v,b}$ for $p \in P^{tr}_{v,b}$. However, when making inference, the responsibility each component should take to "explain" a missing value is unknown, because responsibilities depend on observed target values, according to (12). We have to use Π, the individualized mixing weights, as an approximation of the responsibilities in inference. As defined in (10), Π only depends on the inputs, therefore, we can calculate them when making inferences on the test set.

In a standard mixture model, we could use $\pi^{(k)}_{v,b}$, which is the average of responsibilities of the kth component across all training patients, as a fixed weight that the kth component should contribute to impute missing values $x^{mis}_{:,v,b}$ for all test patients. However, patient time series can be very different and the confidence of predictions by the GP component can vary largely across different patient cases. A fixed weight can not reflect such variation in prediction confidence.

We shall view an individualized mixing weight as an approximation of how much responsibility a component should take to impute the missing value for a particular patient case. It is tailored for each patient. To visualize how individualized mixing weights help to produce better estimates, in Fig. 4, we plot the distribution of the individualized weights Π of the GP component in the training set and compare it with the distribution of the optimized responsibility values Q. The responsibilities that the GP component should take can vary a lot in different patient cases, especially on the synthetic dataset, which implies that it is more reasonable for patients to get individualized mixing weights than a fixed weight. We also observe that the individualized mixing weights reasonably mimic the distribution of the optimized responsibilities on



(a) Chloride, T1, d=0 (b) Chloride, T6, d=0



(c) Chloride, T1, d=0.5 (d) Chloride, T6, d=0.5

Fig. 4. A comparison between individualized mixing weights Π and optimized responsibilities Q that GP component should take to "explain" observed measurements for training patients. The plots are from the real-world MIMIC (d=0) and synthetic MIMIC (d=0.5) dataset, and for the mixture models of Chloride at time point 1 and 6. The distributions of the optimized responsibilities are shown in blue and the distributions of individualized mixing weights are in yellow.

the training set. The improvement of our model on the test set attests that the individualized weights approximate the responsibilities better than fixed weights.

Appendix C
Parameter Estimation in EM

In the E (Expectation) step, we calculate the responsibilities $w_{p,v,b}^{(k)} = Q_p(q_{v,b} = k)$ for $p \in P_{v,b}^{tr}$ using the current values of the parameters in iteration j:

$$
\begin{aligned}
&[C_{p,v,b}^{(1)}]^{(j)} \\
&= [\pi_{v,b}^{(1)}]^{(j)}[D_{p,v,b}^{(1)}]^{(j)}\mathcal{N}(x_{p,v,b}|x_{p,-v,b}[\beta_{v,b}^{(1)}]^{(j)}, [{\sigma_{v,b}^{(1)}}^2]^{(j)}) \\
&[C_{p,v,b}^{(2)}]^{(j)} \\
&= [\pi_{v,b}^{(2)}]^{(j)}[D_{p,v,b}^{(2)}]^{(j)}\mathcal{N}(x_{p,v,b}|x_{p,v,-b}[\beta_{v,b}^{(2)}]^{(j)}, [{\sigma_{v,b}^{(2)}}^2]^{(j)}) \\
&[C_{p,v,b}^{(3)}]^{(j)} \\
&= [\pi_{v,b}^{(3)}]^{(j)}[D_{p,v,b}^{(3)}]^{(j)}\mathcal{N}(x_{p,v,b}|m^{(3)}(\alpha_{p,v,b}^{(j)}), \Sigma^{(3)}(\alpha_{p,v,b}^{(j)})) \\
&[D_{p,v,b}^{(k)}]^{(j)} = \mathcal{N}(V_{p,v,b}|[\boldsymbol{\mu}_{v,b}^{(k)}]^{(j)}, [\boldsymbol{\Sigma}_{v,b}^{(k)}]^{(j)}), k = 1, 2, 3 \\
&[w_{p,v,b}^{(k)}]^{(j)} = \frac{[C_{p,v,b}^{(k)}]^{(j)}}{\sum_{i=1}^{3}[C_{p,v,b}^{(i)}]^{(j)}}, k = 1, 2, 3
\end{aligned}
$$

Let $Z_{v,b} = (x_{p,-v,b})_{p \in P_{v,b}^{tr}}$ and $Y_{v,b} = (x_{p,v,-b})_{p \in P_{v,b}^{tr}}$. In the M (Maximization) step, we re-estimate the parameters in iteration $(j+1)$ using the jth responsibilities:

$$
\begin{aligned}
[\pi_{v,b}^{(k)}]^{(j+1)} &= \frac{1}{|P_{v,b}^{tr}|}\sum_{p \in P_{v,b}^{tr}}[w_{p,v,b}^{(k)}]^{(j)}, k = 1, 2, 3 \\
[\boldsymbol{\mu}_{v,b}^{(k)}]^{(j+1)} &= \frac{\sum_{p \in P_{v,b}^{tr}}[w_{p,v,b}^{(k)}]^{(j)}V_{p,v,b}}{\sum_{p \in P_{v,b}^{tr}}[w_{p,v,b}^{(k)}]^{(j)}} \\
[\boldsymbol{\Sigma}_{v,b}^{(k)}]^{(j+1)} &= \frac{1}{\sum_{p \in P_{v,b}^{tr}}[w_{p,v,b}^{(k)}]^{(j)}}\sum_{p \in P_{v,b}^{tr}}[w_{p,v,b}^{(k)}]^{(j)}[U_{p,v,b}^{(k)}]^{(j+1)} \\
[U_{p,v,b}^{(k)}]^{(j+1)} &= \{V_{p,v,b} - [\boldsymbol{\mu}_{v,b}^{(k)}]^{(j+1)}\}\{V_{p,v,b} - [\boldsymbol{\mu}_{v,b}^{(k)}]^{(j+1)}\}' \\
[\beta_{v,b}^{(1)}]^{(j+1)} &= \{\{Z'_{v,b}[\mathbf{w}_{v,b}^{(1)}]^{(j)}Z_{v,b}\}^{-1}Z'_{v,b}[\mathbf{w}_{v,b}^{(1)}]^{(j)}x_{:,v,b}^{obs}\}' \\
[{\sigma_{v,b}^{(1)}}^2]^{(j+1)} &= \frac{1}{\sum_{p \in P_{v,b}^{tr}}[w_{p,v,b}^{(1)}]^{(j)}}[S_{v,b}^{(1)}]^{(j+1)} \\
[S_{v,b}^{(1)}]^{(j+1)} &= \sum_{p \in P_{v,b}^{tr}}[w_{p,v,b}^{(1)}]^{(j)}\{x_{p,v,b} - x_{p,-v,b}[\beta_{v,b}^{(1)}]^{(j+1)}\}^2 \\
[\beta_{v,b}^{(2)}]^{(j+1)} &= \{\{Y'_{v,b}[\mathbf{w}_{v,b}^{(2)}]^{(j)}Y_{v,b}\}^{-1}Y'_{v,b}[\mathbf{w}_{v,b}^{(2)}]^{(j)}x_{:,v,b}^{obs}\}' \\
[{\sigma_{v,b}^{(2)}}^2]^{(j+1)} &= \frac{1}{\sum_{p \in P_{v,b}^{tr}}[w_{p,v,b}^{(2)}]^{(j)}}[S_{v,b}^{(2)}]^{(j+1)} \\
[S_{v,b}^{(2)}]^{(j+1)} &= \sum_{p \in P_{v,b}^{tr}}[w_{p,v,b}^{(2)}]^{(j)}\{x_{p,v,b} - x_{p,v,-b}[\beta_{v,b}^{(2)}]^{(j+1)}\}^2 \\
[\theta_{v,b}]^{(j+1)} &= G([\mathbf{w}_{v,b}^{(3)}]^{(j)}, [\theta_{v,b}]^{(j)}, x_{:,v,b}^{obs}, Y_{v,b}, t_{:,v,:})
\end{aligned}
$$

where $[\mathbf{w}_{v,b}^{(k)}]^{(j)}$ is the vector of $[w_{p,v,b}^{(k)}]^{(j)}$ for $p \in P_{v,b}^{tr}$ in iteration j. The kernel parameters $\theta_{v,b}$ of GP models are evaluated by function G, a gradient descent method that calculates the estimates of $[\theta_{v,b}]^{(j+1)}$ to maximize $\mathcal{L}_{v,b}(\gamma)$, using $[\theta_{v,b}]^{(j)}$ as the starting point. The first order derivatives of $\mathcal{L}_{v,b}(\gamma)$ with respect to $\theta_{v,b}$ that are used in G are given in Appendix E.

Appendix D
GP Model

We assume the GP model discussed here in a mixture model for a certain variable and time, and thus we exclude the subscripts v and b. We use $x_{p,t}$ to denote a measurement of the time series x_p at time t for patient p of a certain variable. We use $x_{p,-t}$ to denote a time series without the measurement at time t. The GP model is given by

$$
\begin{aligned}
x_{p,t} &= \mu_{p,t} + f(t), \\
f(t) &\sim \mathcal{GP}(0, \mathcal{K}(t, t'))
\end{aligned}
$$

where $\mu_{p,t}$ is the overall mean of the model and $f(t)$ is a Gaussian process with mean of 0 and covariance of $\mathcal{K}(t, t')$. Following the maximum likelihood approach, the best linear unbiased predictor (BLUP) [73] at t and the mean squared error are

$$
\begin{aligned}
m^{(3)}(\theta, x_{p,-t}, \bar{t}) &= (\frac{1 - r^T R^{-1}1_n}{1_n^T R^{-1}1_n}1_n^T + r^T)R^{-1}x_{p,-t} \\
\Sigma^{(3)}(\theta, x_{p,-t}, \bar{t}) &= \sigma_f^2[1 - r^T R^{-1}r + \frac{(1 - 1_n^T R^{-1}r)^2}{1_n R^{-1}1_n}]
\end{aligned}
$$

where $r_t(t') = corr(f(t), f(t'))$, r is the vector of $r_t(t')$ for all possible t, $\bar{t}$ is a vector of time except for time t, R is the $(B-1) \times (B-1)$ correlation matrix and the correlation

function is given by $R_{t,t'} = \exp(-\theta|t-t'|^2)$. The estimator σ^2 is given by

$$\sigma_f^2 = \frac{C^T R^{-1} C}{n}, C = x_{p,-t} - 1_n(1_n^T R^{-1} 1_n)^{-1}(1_n^T R^{-1} x_{p,-t})$$

where 1_n is a vector with length $(B-1)$ of all ones.

Appendix E
Partial Derivatives in GP

To simplify the notations, we assume that the likelihood function L under consideration is for a mixture model for a certain variable and time. The partial derivative with respect to Gaussian process parameters θ is

$$\frac{\partial L}{\partial \theta} = \sum_{p=1}^{|p^{tr}|} w_p \frac{\partial}{\partial \theta} \ln \mathcal{N}(x_{p,t}; m^{(3)}(\theta, x_{p,-t}, \bar{t}), \Sigma^{(3)}(\theta, x_{p,-t}, \bar{t})).$$

Letting $g_p(\theta) = m^{(3)}(\theta, x_{p,-t}, \bar{t})$ and $h_p(\theta) = \Sigma^{(3)}(\theta, x_{p,-t}, \bar{t})$, we have

$$\begin{aligned}
\frac{\partial L}{\partial \theta} &= \sum_{p=1}^{|p^{tr}|} w_p \frac{\partial}{\partial \theta} \ln \mathcal{N}(x_{p,t}; g_p(\theta), h_p(\theta)) \\
&= \sum_{p=1}^{|p^{tr}|} w_p \frac{\partial}{\partial \theta} \{\ln \frac{1}{\sqrt{2\pi h_p(\theta)}} - \frac{[x_{p,t} - g_p(\theta)]^2}{2h_p(\theta)}\} \\
&= \sum_{p=1}^{|p^{tr}|} w_p \{-\frac{1}{2h_p(\theta)} \frac{\partial h_p(\theta)}{\partial \theta} - \frac{\partial}{\partial \theta} \frac{[x_{p,t} - g_p(\theta)]^2}{2h_p(\theta)}\} \\
&= \sum_{p=1}^{|p^{tr}|} w_p \{-\frac{1}{2h_p(\theta)} \frac{\partial h_p(\theta)}{\partial \theta} \\
&- \frac{1}{2h_p^2(\theta)} \{2[x_{p,t} - g_p(\theta)][-\frac{\partial g_p(\theta)}{\partial \theta}] h_p(\theta) \\
&- \frac{\partial h_p(\theta)}{\partial \theta} [x_{p,t} - g_p(\theta)]^2\}\} \\
&= \sum_{p=1}^{|p^{tr}|} w_p \{-\frac{1}{2h_p(\theta)} \frac{\partial h_p(\theta)}{\partial \theta} \\
&+ \frac{[x_{p,t} - g_p(\theta)] \frac{\partial g_p(\theta)}{\partial \theta}}{h_p(\theta)} + \frac{\frac{\partial h_p(\theta)}{\partial \theta} [x_{p,t} - g_p(\theta)]^2}{2h_p^2(\theta)}\}.
\end{aligned}$$

Then $\frac{\partial g_p(\theta)}{\partial \theta}$ and $\frac{\partial h_p(\theta)}{\partial \theta}$ are given by

$$\begin{aligned}
\frac{\partial g_p(\theta)}{\partial \theta} &= (\frac{\partial H_1}{\partial \theta} R^{-1} + H_1 \frac{\partial R^{-1}}{\partial \theta}) x_{p,-t} \\
\frac{\partial h_p(\theta)}{\partial \theta} &= \sigma_f^2 \frac{\partial H_3}{\partial \theta} + \frac{\partial \sigma_f^2}{\partial \theta} H_3
\end{aligned}$$

where H_1, $\frac{\partial H_1}{\partial \theta}$, H_3 and $\frac{\partial H_3}{\partial \theta}$ are given as follows:

$$\begin{aligned}
H_1 &= \frac{[1 - (rR^{-1}1_n)]}{1_n^T R^{-1} 1_n} 1_n^T + r \\
\frac{\partial H_1}{\partial \theta} &= \frac{-(\frac{\partial r}{\partial \theta} R^{-1} + r \frac{\partial R^{-1}}{\partial \theta}) 1_n (1_n^T R^{-1} 1_n)}{1_n^T R^{-1} 1_n^2} \\
&- \frac{(1_n^T \frac{\partial R^{-1}}{\partial \theta} 1_n)[1 - (rR^{-1}1_n)]}{1_n^T R^{-1} 1_n^2} 1_n^T + \frac{\partial r}{\partial \theta}
\end{aligned}$$

$$\begin{aligned}
fc &= (1 - 1_n^T R^{-1} r^T)^2 \\
gc &= 1_n^T R^{-1} 1_n \\
\frac{\partial fc}{\partial \theta} &= 2(1 - 1_n^T R^{-1} r^T)[-1_n^T (\frac{\partial R^{-1}}{\partial \theta} r^T + R^{-1} \frac{\partial r^T}{\partial \theta})] \\
\frac{\partial gc}{\partial \theta} &= 1_n^T \frac{\partial R^{-1}}{\partial \theta} 1_n \\
H_2 &= \frac{(1 - 1_n^T R^{-1} r^T)^2}{1_n^T R^{-1} 1_n} \\
\frac{\partial H_2}{\partial \theta} &= \frac{\frac{\partial fc}{\partial \theta} gc - \frac{\partial gc}{\partial \theta} fc}{gc^2}
\end{aligned}$$

$$\begin{aligned}
H_3 &= 1 - (rR^{-1}r^T) + H_2 \\
\frac{\partial H_3}{\partial \theta} &= -(\frac{\partial r}{\partial \theta} R^{-1} r^T + r \frac{\partial R^{-1}}{\partial \theta} r^T + r R^{-1} \frac{\partial r^T}{\partial \theta}) + \frac{\partial H_2}{\partial \theta} \\
H_4 &= x_{p,-t} - 1_n \frac{(1_n^T R^{-1} x_{p,-t})}{1_n^T R^{-1} 1_n} \\
\frac{\partial H_4}{\partial \theta} &= -1_n \frac{1}{(1_n^T R^{-1} 1_n)^2} [(1_n^T \frac{\partial R^{-1}}{\partial \theta} x_{p,-t})(1_n^T R^{-1} 1_n) \\
&- (1_n^T \frac{\partial R^{-1}}{\partial \theta} 1_n)(1_n^T R^{-1} x_{p,-t})] \\
\frac{\partial \sigma_f^2}{\partial \theta} &= \frac{1}{n} [(\frac{\partial H_4}{\partial \theta})^T R^{-1} H_4 + H_4^T \frac{\partial R^{-1}}{\partial \theta} H_4 + H_4^T R^{-1} \frac{\partial H_4}{\partial \theta}].
\end{aligned}$$