A Non-parametric Conditional Factor Regression Model for Multi-Dimensional Input and Response

Ava Bargi\textsuperscript{a} Richard Yi Da Xu\textsuperscript{b} Zoubin Ghahramani Massimo Piccardi\textsuperscript{c}
University of Technology, Sydney\textsuperscript{a,b,c}
University of Cambridge

Abstract

In this paper, we propose a non-parametric conditional factor regression (NCFR) model for domains with multi-dimensional input and response. NCFR enhances linear regression in two ways: a) introducing low-dimensional latent factors leading to dimensionality reduction and b) integrating the Indian Buffet Process as prior for the latent layer to dynamically derive an optimal number of sparse factors. Thanks to IBP’s enhancements to the latent factors, NCFR can significantly avoid over-fitting even in the case of a very small sample size compared to the dimensionality. Experimental results on three diverse datasets comparing NCFR to a few baseline alternatives give evidence of its robust learning, remarkable predictive performance, good mixing and computational efficiency.

1 Introduction

With the exponential growth in data generation, multi-variate problems with high-dimensional input and output are becoming more common. Volatility matrix estimation and price forecasting in finance (\textsuperscript{?}), as well as action prediction, view-to-view recognition (\textsuperscript{?}) and pose estimation (\textsuperscript{?}) in computer vision are examples of such problems. Linear regression has been a long-standing, simple yet effective prediction tool for many domains. However, most of the existing solutions cater for uncorrelated one-dimensional response, leaving a visible gap for multi-variate cases. In this paper, we propose an elaborate regression model to cater for multi-variate prediction contexts, while modelling the correlations among response variables. Empirical results on three diverse datasets prove significant accuracy and good mixing. The proposed model also exhibits considerable robustness, regardless of diverse ratios between the number of available samples and dimensions across the datasets.

Let us begin with the multi-dimensional classic regression model $Y = RX + E$, where $Y$ is the $D_q$-dimensional response over $N$ observations ($D_q \times N$) and $X$ is the $D_p \times N$ input, regressed by $R$ and added with diagonal Gaussian noise, $E$. For large $D_p$ and $D_q$, $R$ would be a large matrix, likely to overfit and imposing matrix multiplications of order $O(D_q D_p N)$, that are computationally costly.

As a step forward, we can improve the model by introducing a latent factor, $Z(K \times N)$, not only bridging $X$ and $Y$, but jointly reducing their ranks to $K \ll D_p, D_q$: $Y = QZ + E_y, Z = PX + E_z$. Such latent factors also improve the noise model by decoupling it for input and response into separate $E_z$ and $E_y$ (\textsuperscript{?}). Let us call this parametric model conditional factor regression (CFR). Variants of such models are studied in (\textsuperscript{?}), (\textsuperscript{?}), (\textsuperscript{?}) and (\textsuperscript{?}), all being subject to a prior decision making on the optimal latent dimensionality ($K$) through trial and error or domain knowledge.

To improve on this, we propose a ‘non-parametric’ conditional factor regression (NCFR) model: a novel Bayesian non-parametric treatment to multi-variate linear regression, enhanced by (a) introducing latent factors with integrated dimensionality reduction mechanisms and (b) finding the optimal reduced dimensions ($K$) by exploiting an Indian Buffet Process (IBP) prior (\textsuperscript{?}), thereby avoiding overfitting through sparse latent factors. The marriage of these properties makes the resulting NCFR resilient to noise and effective in the presence of limited sample size in real-life problems. Empirical results on three diverse datasets give evidence to this claim. In section 2, we explore the background research on similar models, followed by an articulate description of NCFR model parameters and inference in sections 3 and 4. Through experi-
ments in section 5, we evaluate and compare the above-mentioned models, following with the Conclusion.

2 Background

There are numerous studies in the regression and Bayesian non-parametric literatures. Here, we tend to confine the focus on latent factor analysis/regression models and similar alternatives, concluding with a brief summary of the Indian Buffet Process. The notion of latent factor regression has been introduced with varying terminology and design. The Bayesian factor regression model (\cite{1}) represents a regression model particularly suited for large input size and small observation number. It uses a latent factor between input and response, assuming them both dependent on the latent factor. On the contrary, the spectral latent variable model proposed in (\cite{2}) shares the same design as our model, creating a transitive dependency of response over the latent factor, in turn conditioned over the input (fig. \cite{3}). The dependencies between high-dimensional inputs and responses are channelled via a low-dimensional latent manifold using mixtures of Relevance Vector Machines (RVM).

Both the above models require a parametric choice of \( K \) (dimensionality of the latent layer). Caron and Doucet have remedied this through proposing a class of priors based on scale mixture of Gaussians over the regressor (\cite{4}). Their sparse Bayesian non-parametric model is essentially an infinite sparse regressor, correlating an infinite dimensional input to a single dimensional response through Levy processes. This solution, however, is not computationally efficient for high-dimensional regression. In the context of factor analysis, \( \beta \) use a multiplicative Gamma process shrinkage prior on the factor loading matrix to permit infinite number of factors. Similarly, \( \gamma \) induce the basis selection by choosing a shrinkage prior that allows many of the loadings to be close to zero. Alternatively, one can utilise an IBP prior to select the relevant factors for each observation. Infinite sparse factor models exploit this property of IBP to enhance factor analysis, through sparsifying either the factors (\cite{5}) or the factor loading matrix (\cite{6}). The choice of which parameter to sparsify depends on the problem domain and cannot be used interchangeably. In cases where each response variable is identically applied to all samples, the latter approach is used. However, in our model each sample has its own sparse latent factors common to all dimensions. Hence we impose sparsity over the latent factors (\( Z \)), yet cascading conditional dependence of \( Y \) on \( Z \) and transitively over \( X \).

2.1 Indian Buffet Process: the binary mask

The art of an Indian buffet process is that of constructing an infinite sparse binary matrix, \( S \), with features in the rows and samples as columns. Griffith and Ghahramani (2006) initially introduce a finite case for \( S \) consisting of \( K \) features and further extend it to the asymptotic infinite version where \( K \rightarrow \infty \). Each feature \( k \) is active with a Binomial likelihood parametrised by \( \pi_k \), integrated out for convenience. The resulting density for \( S \) is a normalised Poisson distribution:

\[
P(S|\alpha) = \frac{\alpha^K}{\prod_{h>0} K_h!} \exp(-\alpha H_N) \prod_{k=1}^{K} \frac{(N - m_k)!(m_k - 1)!}{N!}
\]

where \( \alpha \) is the strength parameter and \( m_k = \sum_{n=1}^{N} s_{k,n} \) is the number of data points for which the \( k \)th feature is active. \( H_N = \sum_{j=1}^{N} \frac{1}{j} \) is the \( N \)-th harmonic number and \( K_h \) is the number of rows in which the numerical value of the sequence of binary digits is equivalent to the decimal number \( h \).

Using the above distribution, the Indian buffet analogy is phrased as a buffet with an infinite number of dishes allowing customers to try unlimited options. The probability of choosing each dish is driven by previous customers’ choices; with the exception of untried dishes, the number of which is drawn from a Poisson prior. In other words, each customer \( i \) will select an already tried dish with probability \( \frac{m_k}{i} \), and decide to try a new dish with probability \( \text{Poisson}(\frac{h}{i}) \). Thanks to exchangeability, each customer can be treated as the last; thus, informing each decision with all other choices (\( s_{-ki} \)):

\[
P(s_{ki} = 1|s_{-ki}) = \frac{m_{k,-i}}{N}, \quad m_{k,-i} = \sum_{n \neq i} s_{kn}
\]

3 The model

The proposed Non-parametric Conditional Factor Regression (NCFR) consists of two linear Gaussian transforms, linked through a sparse latent layer. To maintain reasonable generality, we have assumed diagonal covariance matrices for the noise, reserving the option for isotropic variances in simpler models. As mentioned earlier, an important challenge remains that of selecting the optimal dimensionality for the latent factor, yielding the best regression and dimensionality reduction performance. In conventional cases, this is a rigid decision. However, by adding an IBP prior to the model, we have derived an infinite sparse \( Z \) to best fit the data. Please note that the term non-parametric in the title NCFR is merely to indicate the usage of
4 Inference

Given the input and response observations $X$ and $Y$, we infer model’s random parameters jointly in a posterior probability, using Gibbs sampling. Additional Metropolis-Hastings (MH) steps are utilised particularly for deriving the newly activated features in $S$ (5). This section briefly introduces the inference steps for all random variables in Equation (5). The NCFR inference algorithm is shown in Algorithm 2.

3The model could be considered semi-parametric in that it combines a linear regression structure with a nonparametric process for selecting the sparsity in the parameter space.
cludes the parameter of interest \( z_{kn} \), whereas the latter (residue term) excludes \( z_{kn} \) through a reduced product \( [Q_{zn}]_{z_{kn}=0} \) with one fewer element in \( z_{n} \). Yet, to maintain the dimensional compatibility, we have kept the \( k \)-th element and made it equal to zero. Hence, the marginal likelihood for the active case is distributed as in Equation ???. As can be seen in \( N_i \), the conditional probability \( P(y_n|x_n) \) models the correlations among response variables \( (y_{kn} \text{ to } y_{Dn}) \) thanks to the full covariance matrix, \( \Psi_y + q_k \sigma_z q_z^{-1} q_k^T \). In case of inactive \( s_{kn} \), the Hadamard product \( s_{kn} \odot z_{kn} \) is inevitably zero for every \( z_{kn} \). Therefore, the prior probability for marginalisation is no longer informative, resulting in the \( D_l \) below. The likelihood ratio \( (r_l) \) can be obtained from the \( N_i / D_l \) ratio.

\[
\begin{align*}
N_i &= \int N(y_n|Q_{zn}, \Psi_y)N(z_{kn}|p_k x_n, \sigma_{zk}) dz_{kn} \\
&= N(y_n|q_k p_k x_n + [Q_{zn}]_{z_{kn}=0}, \Psi_y + q_k \sigma_z q_z^{-1} q_k^T) \quad (6) \\
D_l &= N(y_n|[Q_{zn}]_{z_{kn}=0}, \Psi_y)
\end{align*}
\]

The prior ratio \( (r_p) \) is derived below, following Equation ???. Note that the \( i \)-th observation is not counted for in the ratio, which justifies the \( -1 \) in the denominator. Having derived \( r_l \) and \( r_p \) (eqs. ?? and ??), the posterior ratio for existing features may be derived as a uniform draw over \( \frac{r}{r + 1} \).

\[
r_p = \frac{m_{k, -i}}{N - 1 - m_{k, -i}} \quad (7)
\]

Next, we should decide the number of new features \( (\kappa_n) \) added for the current observation. IBP implies that \( \kappa_n \) is a priori distributed as \( \text{Poisson} \left( \frac{\alpha}{N} \right) \). Yet, there have been different approaches for sampling \( \kappa_n \) from its posterior. In (??) \( \kappa_n \) is sampled through a MAP estimate of various values for \( \kappa_n \), ranging from zero to an upper bound. However, (??) proposes a Metropolis-Hastings step to some random \( \kappa_n \) and evaluates the acceptance through posterior ratios. Each MH jump from \( \xi \rightarrow \xi^* \) is done with a probability \( J(\xi^* | \xi) \) with varying underlying assumptions. A basic approach considers the prior on \( \xi^* \) as candidate function, i.e. \( J(\xi^* | \xi) = P(\xi)P(\xi^*) \).

\[
r_{\xi \rightarrow \xi^*} = \frac{P(\xi^* | y_n, \ldots, P(\xi)P(\xi^*)}{P(\xi | y_n, \ldots, P(\xi)P(\xi^*)} = \frac{N(y_n|q_k^T p_k^T x_n + [Q_{zn}]_{z_{kn}=0}, \Psi_y + q_k \sigma_z q_z^{-1} q_k^T)}{N(y_n|[Q_{zn}]_{z_{kn}=0}, \Psi_y)} \quad (8)
\]

In the equations above, \( _* \) denotes the remaining parameters and \( * \) superscripts indicate extended parameters with \( \kappa_n \) new features. The jump is ultimately accepted with probability \( \min(1, r_{\xi \rightarrow \xi^*}) \). In order to add \( \kappa_n \) new features, we need to add an equal number of new columns and rows to \( Q \) and \( P \), respectively. These new vectors are denoted with prime notations as in \( \xi^* = \{ \kappa_n, q', p' \} \). Please note that along with activating \( \kappa_n \) new features, an equivalent number of weight elements \( (z') \) is needed. Such weights are marginalised for simplicity.

### 4.2 Factor loading matrices and latent factor weights

Inference of factor loading matrices can be uniformly described for new and existing features, through sampling \( K \) independent vectors in \( Q \) and \( P \). Following the graphical model (fig. ??), the factor loading matrices are independent from each other through \( Z \). Thus, their posterior probabilities can be independently derived. We thus infer column vectors \( q_k \) through sampling individual elements, \( P(q_k|Y, Q^{-k}, Z, \Psi_y, \Psi_q) = \prod_{i=1}^{D_k} N(q_{ik|\mu_{ik, \Sigma_{ik}}}, \forall q_{ik} : i \in [1, D_k] \). Following a similar approach, we derive the posterior for \( p_k \).

\[
\begin{align*}
N(q_{ik}|\mu_{q_{ik}, \sigma_{q_{ik}}}) \propto N(y_i|q_i z_k + [q_i|Z]_{z_{ik}=0}, \sigma_{yi})N(q_{ik}|0, \sigma_{yi}) : \\
\sigma_{q_{ik}} = \frac{\sigma_{yi} \sigma_{yi}}{\sigma_{yi} + \sigma_{zi} z_k^T z_k} \mu_{q_{ik}} = \frac{\sigma_{yi}(y_i - [q_i|Z]_{z_{ik}=0}) z_k^T}{\sigma_{yi} + \sigma_{zi} z_k^T z_k} \\
N(p_k|\mu_{p_k}, \Sigma_{p_k}) \propto N(z_k|p_k X, \sigma_{zk})N(p_k|0, \sigma_{p_k} I_{D_p}) : \\
\Sigma_{p_k} = \sigma_{p_k} I_{D_p} + \sigma_{zk} (XX^T)^{-1}, \mu_{p_k} = \Sigma_{p_k} X z_k^T / \sigma_{zk}
\end{align*}
\]

Each active feature \( s_{kn} \) is assigned a weight \( z_{kn} \), the element-wise product of which forms the rank-deficient sparse latent layer between the high dimensional input \( X \) and response \( Y \). The posterior on each \( z_{kn} \) can be derived as follows.

\[
\begin{align*}
N(z_{kn}|\mu_{z_{kn}, \Sigma_{z_{kn}}}) \propto N(y_n|q_k z_{kn} + [Q_{zn}]_{z_{kn}=0}, \Psi_y)N(z_{kn}|p_k x_n, \sigma_{zk}) : \\
\sigma_{z_{kn}} = \left( \frac{1}{\sigma_{zk}} + q_k^T \Psi_y^{-1} q_k \right)^{-1}, \\
\mu_{z_{kn}} = \sigma_{z_{kn}} q_k^T \Psi_y^{-1} (y_n - [Q_{zn}]_{z_{kn}=0} + p_k x_n) / \sigma_{zk}
\end{align*}
\]

### 4.3 Noise and factor loading covariances and IBP parameters

NCFR adopts diagonal noise models, \( \Psi_y \) and \( \Psi_z \). The factor loading covariances, \( \Psi_q \) and \( \Psi_p \), are also diagonal. Following (??), we sample elements on the main
diagonal of the above covariance matrices through Inverse Gamma priors. For simpler models the diagonal variances can be simplified into an isotropic model. Finally, IBP parameter $\alpha$ is sampled through a conjugate Gamma prior, as follows. The distribution for $P(S|\alpha)$ and other notations are introduced in section 7.

$$
IG(\sigma_{\nu}|a + \frac{N}{2}, b + tr(E_{\nu}^TE_{\nu})) \propto N(E_{\nu}|0, \sigma_{\nu})IG(\sigma_{\nu}|a, b)
$$

$$
IG(\sigma_{zk}|a + \frac{m_k}{2}, b + tr(E_{zk}^TE_{zk})) \propto N(E_{zk}|0, \sigma_{zk})IG(\sigma_{zk}|a, b)
$$

$$
IG(\sigma_{qi}|c + \frac{D_q}{2}, d + tr(q_kq_k^T)) \propto N(q_k|0, \sigma_{qi})IG(\sigma_{qi}|c, d)
$$

$$
IG(\sigma_{pj}|c + \frac{D_p}{2}, d + tr(p_k p_k^T)) \propto N(p_k|0, \sigma_{pj})IG(\sigma_{pj}|c, d)
$$

$$
G(\alpha|K + e, f + H_N) \propto Poisson(S|\alpha)G(\alpha|e, f)
$$

5 Experiments

As an attempt to evaluate the models above, we have synthetically generated $X$ and $Y$ through an IBP-like process, since following the exact IBP steps would create bias towards the model. The synthetic data use a 5D latent layer to link a 70D input and 50D response, which are realistic values in computer vision problems (see for instance?). $Z$ is calculated as the product of a standard Gaussian random $X$ and zero-mean diagonal Gaussian $P$ with added Gaussian noise. It is then masked by a binomially-distributed $S$. The resulting sparse $Z$ is utilised to generate $Y$, along with $Q$ and diagonal Gaussian noise $\Psi_q$ (eq. ??). We train the models with 70 percent of observations and use the remaining unseen data for test. The trained parameters ($Q$ and $P$) are obtained from the sample with highest likelihood amongst the last 100. Hence, prediction can be efficiently performed using a single linear transformation $Y = QPX$. The prediction results ($e = || y_n - \tilde{y}_n || / || y_n ||$) are reported in Table ??, showing noticeably lower errors in the generic and isotropic NCFR. Removing the IBP prior, CFR variants with diagonal noise perform better than the EM alternatives with unit spherical variance. Yet, the classic linear regression model is largely inefficient due to drastic overfitting and error. We further visualise
Table 1: Models and variants exploited and compared in the experiments. Variants are denoted with their italic abbreviations.

| Classic Linear Regression | Conditional Factor Regression | Non-parametric CFR |
|---------------------------|-------------------------------|--------------------|
| Full-rank linear regression (FRR) | EM, large $K$ (EMCFR) | Generic (Gen) |
| EM, medium $K$ (EMCFRm) | EM, small $K$ (EMCFRs) | Fixed $\alpha$ (Fix$\alpha$) |
| MCMC, large $K$ (CFRl) | MCMC, medium $K$ (CFRm) | Isotropic noise (IsoN) |
| MCMC, small $K$ (CFRs) | | Isotropic loading (IsoL) |

Figure 2: IsoN NCFR performance on synthetic data: (a) Heat map for visual comparison of real $Y_{D_q \times N_{\text{test}}}$ matrix vs. predicted $Y$. (b) Scatter plot of real $Y$ vs. predicted $Y$ with all dimensions superimposed on the same plot. The plot includes $D_q \times N_{\text{test}}$ points all aligned around $y = x$, depicting a significant predictive performance over all dimensions. The figure is best viewed in colour.

Figure 2 illustrates the accuracy in IsoN variant via heat maps and scatter plots of predicted $Y$ vs. real $Y$ (fig. ??(a)). The scatter plot also shows significant accuracy, aligning the points along the line $y = x$ (fig. ??(b)). NCFR tends to slightly overfit, converging to larger $K$ than the original dimensionality of the latent layer. The IsoN NCFR reported above converges to 7 active features. However, we found that the active features were clearly clustered into two groups in terms of $S$ sparsity rates: 5 fully dense features and the other 2 highly sparse. These results along with the accuracy performance reported above prove that IsoN and Gen NCFR can efficiently fit the synthetic data.

5.2 Human pose prediction

We have used the TUM Assistive Kitchen dataset (?), containing motion capture sequences of 9 complex everyday activities in a kitchen with the common scenario of collecting utensils from cupboards and laying a table. Each frame is described by an 81D feature vector of body joint coordinates. Our target is to predict the next frame’s pose given the current pose within an activity in an auto-regressive manner, maximising the predictive probability $P(y_n|y_{n-1}, Q, P, \Psi_y, \Psi_z)$. In other words, we regress $y_n$ against $x_n = y_{n-1}$ preserving independence through the Markov assumption. A possible use of this posterior is to avoid the costly task of pose data extraction in every frame, by computing them for only a few samples and predicting the rest with minimal computation. We have trained the models with 6 occurrences of the action “Carrying while walking” and tested it over an unseen occurrence, using sequences performed by Subject1: 1-1 to 1-5.

Figure ??(a) illustrates the normalised prediction error for the generic sequence 1-2 across the models. All variants of our proposed model exhibit an average error of approximately 7%, slightly favoring isotropic noise and factor loading variance over the generic model. Utilizing IBP in NCFR has noticeably leveraged performance and model selection compared to the next best models (MCMC-CFR variants), thanks to sparsity and adaptive inference of the optimal dimensionality in the latent layer. Full-rank linear regression and the EM-CFR with unit variance produced considerably more error due to overfitting and inefficient noise modeling. The performance trend is roughly similar across different sequences with a minimum average error of 4% for sequence 1-3. It is worth noting that action dynamics are occasionally very non-linear, mak-
Table 2: Prediction error percentage ($\epsilon = 100 \times \frac{\| y_n - \tilde{y}_n \|}{\| y_n \|}$) for synthetic prediction.

|                      | Classic Linear Regression | Conditional Factor Regression | Non-parametric CFR |
|----------------------|---------------------------|-----------------------------|-------------------|
| FRR                  | 370.52 ± 259.34           | 6.68 ± 2.90                 | 6.50 ± 2.83       |
| CFR                 |                           | 6.64 ± 2.90                 |                   |
| CFRm                |                           | 89.23 ± 14.57               |                   |
| CFRs                |                           |                             |                   |
| Gen                 |                           |                             |                   |
| Fixx                |                           |                             |                   |
| IsoN                |                           | 6.60 ± 2.76                 |                   |
| IsoL                |                           | 6.73 ± 2.88                 |                   |

Figure 3: Boxplot of prediction performance on TUM kitchen data, sequence 1-2: (a) Normalised prediction errors are presented over 148 frames of $Y$ and sorted across the 3 categories of models. The top 6 variants are magnified in the plot on the right (b), showing noticeably better performance for NCFR variants appearing on the first 3 boxes.

5.3 Natural gas consumption prediction

We have also tested the proposed models in a considerably different domain: energy supply/consumption forecast. The US Energy Information Administration (?) publishes reports on various sources of energy, including natural gas. We have exploited 25 variables on natural gas monthly reports on supply, consumption and inventories in 2009-2012 (48 months). Our target is to forecast the supply and consumption in 2011-12 given the data for 2009-10, each including 24 samples. Natural gas usage exhibits seasonal trends, allowing us to forecast each month independently, according to the respective month sample in previous years. The shapes of $X_{(25 \times 24)}$ and $Y_{(25 \times 24)}$ matrices for this dataset are considerably square-like and smaller compared to the previous two, as the number of samples is even less than the dimensions. Prediction in such contexts is considerably more challenging, due to high risks of overfitting (?). The forecast results are reported in Table ??, achieving the lowest error in NCFR with isotropic noise ($IsoN$). In agreement with the previous experiments, the isotropic variants of NCFR perform the best, followed by the simpler MCMC-CFR models. Due to the challenges of this dataset the MCMC sampler has taken a longer burn-in process (totally 3000 iterations, as opposed to 500 iterations with the synthetic data and 1000 for the kitchen experiment) to ensure convergence. Yet, the small number of samples permits much faster training, as reported in section ??.

Computing the covariance on $Y$ and its eigenvalues, we have attained a rough estimate of the significant dimensions in the data. According to the eigenvalues, there are 7 main factors with different levels of significance (4 strong and 3 weak). NCFR has activated 8 dimensions for the latent layer with high sparsity rate of around 60%.

5.4 Sampling efficiency and computational costs

We next examine the Gibbs sampler’s mixing rate and execution time for the experiments on real data, kitchen and gas. Figure ??(a) visualises the log likelihood mixing for gas dataset, in the last 200 iterations. Since all the sampled variables are utilised in log likelihood calculation, the well mixed results indicate general mixing efficiency in the model. The other two experiments exhibit very similar mixing patterns, not visualised due to lack of space.

Finally, we consider the computational costs to explore
Table 3: Prediction error percentage ($\epsilon = 100 \times \| y_n - \tilde{y}_n \| / \| y_n \|$) for gas consumption forecast.

| Classic Linear Regression | Conditional Factor Regression | Non-parametric CFR |
|---------------------------|-------------------------------|-------------------|
| FRR                       | CFRl                          | Gen               |
| 83.96 ± 9.51              | 19.33 ± 2.28                 | 21.44 ± 1.53      |
|                           | CFRm                          | Fizz              |
|                           | 18.57 ± 2.49                 | 21.00 ± 0.90      |
|                           | CFRs                          | IsoN              |
|                           | 16.51 ± 2.87                 | 10.68 ± 2.65      |
|                           |                               | IsoL              |
|                           |                               | 17.70 ± 4.58      |

Figure 4: Sampling efficiency and computational cost: (a) Log likelihood plot of NCFR gas results with the last 200 iterations magnified, showing convergence and well mixing. (b) Inference time per iteration for TUM kitchen pose prediction (c) Inference time per iteration per active feature for the same kitchen experiment. Please note that CPU time has been divided by $K \times S_{density}$ to provide a precise metric with respect to number of active features and their average sparsity. Combined analysis of (b) and (c) creates a thorough understanding of inference speed.

viability. Figures ??(b, c) illustrate CPU time in the kitchen experiment, run on a basic machine with an Intel i5 (3.10 GHz) processor and 4GB memory. Exploring CPU time per iteration allows us to compare the overall inference time, given 1000 iterations for kitchen data and 3000 for gas. Among the variants with best predictive accuracy, the generic and isotropic NCFR perform more efficiently, while MCMC-CFR variants show a clear correlation between the size of latent layer ($K$) and elapsed time. In order to audit computation time independently from $K$, we have further divided CPU time per iteration by an average metric for active features’ density ($K \times S_{density}$). Removing the impact of $K$, the top NCFR variants (IsoN and Gen) perform equally efficiently as MCMC-CFR, yet providing sparse and adaptive modeling and considerably better accuracy. Experiments on gas data run 100 times faster on average, yet exhibiting very similar trends amongst the variants. The considerable speed compared to the kitchen dataset is attributed to the lower number of samples and dimensions. It is important to acknowledge that FFR and the EM-CFR variants are much faster than NCFR and MCMC-CFR variants since their models are simpler. However, their simplicity, poor noise model and strict linearity result in dramatic overfitting and poor performance.

6 Conclusion

In this paper we have proposed a novel solution for multivariate factor regression. The proposed model offers two improvements over classic linear regression. Through exploiting a latent space of lower dimensionality, NCFR reduces the degrees of freedom and mollifies overfitting. By integrating an IBP prior in the latent factor inference, it generates factors which are both sparse and adaptive in number. Experimental results on a synthetic model and two real datasets prove that NCFR achieves remarkable prediction accuracy while maintaining acceptable computational costs. More importantly, despite the diverse attributes of the three datasets in terms of sample size per dimension, NCFR shows significant resilience and adaptability against noise and overfitting.