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

Ava Bargi[a] **Richard Yi Da Xu**[a] **Zoubin Ghahramani**[b] **Massimo Piccardi**[a]
University of Technology, Sydney[a] University of Cambridge[b]

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 NCRF 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 (Wang and Zou, 2010), as well as action prediction, view-to-view recognition (Kusakunniran et al., 2010) and pose estimation (Bo and Sminchisescu, 2009) in computer vision are examples of such problems. Linear regression has been a long-standing, simple yet effective prediction tool for many domains. However, many of the existing solutions focus on 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 (West, 2003). Let us call this parametric model *conditional factor regression* (CFR). Variants of such models are studied in (West, 2003), (Carvalho et al., 2008), (Teh et al., 2005) and (Bo and Sminchisescu, 2009), 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 (Griffiths and Ghahramani, 2006), 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.

Appearing in Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2014, Reykjavik, Iceland. JMLR: W&CP volume 33. Copyright 2014 by the authors.

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 experiments 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 (West, 2003) 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 (Bo and Sminchisescu, 2009) shares the same design as our model, creating a transitive dependency of response over the latent factor, in turn conditioned over the input (fig. 1). 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 (Caron and Doucet, 2008). 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, Bhattacharya and Dunson (2011) use a multiplicative Gamma process shrinkage prior on the factor loading matrix to permit infinite number of factors. Similarly, Montagna et al. (2012) 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 (Knowles and Ghahramani, 2007) or the factor loading matrix (Knowles and Ghahramani, 2011), (Rai and Daumé III, 2009). 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 $K \to \infty$. Each feature k is active with a Binomial likelihood parametrised by π_k, integrated out for convenience. The resulting density for S is a normalised Poisson distribution:

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

where α is the strength parameter and $m_k = \sum_{n=1}^{N} s_{k,n}$ is the number of data points for which the kth 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 $Poisson(\frac{\alpha}{i})$. Thanks to exchangeability, each customer i could be treated as the last; thus, informing each decision with *all* other choices (s_{-ki}):

$$P(s_{ki}=1|s_{-ki}) = \frac{m_{k,-i}}{N}, \qquad m_{k,-i} = \sum_{n \neq i} s_{kn} \quad (2)$$

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 a non-parametric Bayesian prior, IBP[1]. The model is defined as follows:

$$\begin{aligned} p(Y|Z) &= \mathcal{N}(Y|Q(S \odot Z), \Psi_y) \\ p(Z|X) &= \mathcal{N}(Z|PX, \Psi_z) \quad , \end{aligned} \tag{3}$$

assuming X, Y and Z consist of N independent observations. S is the IBP binary mask applied to Z, determining whether or not each dimension is active for each sample. The weight of each active feature in S is specified via its respective entry in Z.

Ultimately, each of the D_p-dimensional input vectors, x_n in $X_{D_p \times N}$, is regressed into a respective y_n response vector of D_q dimensions, via a sparse K-dimensional z_n that is masked by a relevant binary s_n through element-wise Hadamard product $(s_n \odot z_n)_{K \times N}$. The result is added with Gaussian noise terms, ϵ_{yn} and ϵ_{zn}, with diagonal covariances Ψ_y and Ψ_z. The noise terms collectively form $E_{y(D_q \times N)}$ and $E_{z(K \times N)}$. Q and P are factor loading matrices, comprised of independent Gaussian vectors, $Q = \{q_{:k}\}_{k=1..K}$ and $P = \{p_{k:}\}_{k=1..K}$, with diagonal covariances Ψ_q and Ψ_p. $q_{:k}$ and $p_{k:}$ are later noted as q_k and p_k for simplicity. The individual variances on the diagonal of the above covariance matrices are noted as σ with the relevant subscripts and indices, for instance $\Psi_y = diag(\sigma_{y_1}, ..., \sigma_{y_{D_q}})$. The binary mask matrix S is sampled from an IBP prior with Gamma-distributed α. Figure 1 illustrates the proposed graphical model, the priors of which are defined below.

$$\begin{aligned} &\epsilon_y \sim \mathcal{N}(0, \Psi_y), && \sigma_{yi} \sim IG(a, b), i = 1..D_q \\ &\epsilon_z \sim \mathcal{N}(0, \Psi_z), && \sigma_{zk} \sim IG(a, b), k = 1..K \\ &p_{k:} \sim \mathcal{N}(0, \Psi_p), && \sigma_{pj} \sim IG(c, d), j = 1..D_p \\ &q_{:k} \sim \mathcal{N}(0, \Psi_q), && \sigma_{qi} \sim IG(c, d), i = 1..D_q \\ &S \sim IBP(\alpha), && \alpha \sim G(e, f) \end{aligned} \tag{4}$$

4 Inference

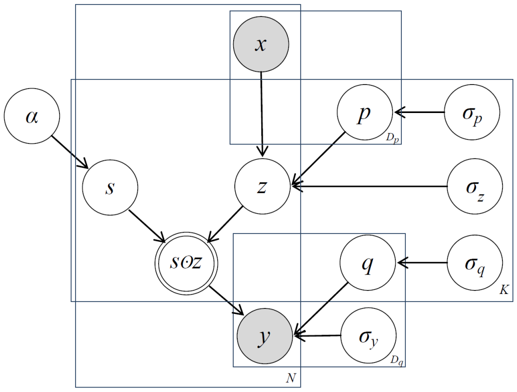

Figure 1: NCFR Graphical model. The node labeled as $s \odot z$ is deterministic, illustrated with double-ring oval notation (Koller and Friedman, 2009). The plate notation is used to clarify the dimensions of matrices introduced with capital letters in the model.

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 (Knowles and Ghahramani, 2007). This section briefly introduces the inference steps for all random variables in Equation 4. The NCFR inference algorithm is shown in Algorithm 1.

4.1 Binary mask

Complying with the definition of IBP and similar infinite processes, $S_{K \times N}$ is a matrix with an infinite number of sparse rows, $s_{k:}$. For simplicity and memory efficiency, we only consider the active features ($\forall\, k \mid \exists s_{kn} = 1$) and sample their related weights. Thus, S could extend or shrink in each iteration, due to customers' choices. In this section, we begin with sampling the existing active features and proceed with adding new ones. To sample the binary elements of S, we form the ratio of posteriors for active vs inactive s_{kn} elements. Thanks to this technique, normalization factors are cancelled out and a uniform random draw can determine whether or not s_{kn} is active. The posterior ratio is decomposed into separate ratios for likelihood and prior (r_l and r_p), as in Equation 5.

$$\begin{aligned} r &= \frac{P(s_{kn} = 1|y_n, s_{-kn}, z_{-kn}, Q, \Psi_y)}{P(s_{kn} = 0|y_n, s_{-kn}, z_{-kn}, Q, \Psi_y)}, \\ r_l &= \frac{N_l}{D_l} = \frac{P(y_n|s_{kn} = 1, s_{-kn}, z_{-kn}, Q, \Psi_y)}{P(y_n|s_{kn} = 0, s_{-kn}, z_{-kn}, Q, \Psi_y)}, \\ r_p &= \frac{P(s_{kn} = 1|s_{-kn})}{P(s_{kn} = 0|s_{-kn})}, \qquad r = r_l.r_p \end{aligned} \tag{5}$$

Before attempting to derive this ratio, the likelihood terms are marginalised with respect to z_{kn}, the latent factor weight of the k-th feature in the n-th observation. However, such weights are later sampled

[1] The 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

for all active s_{kn}. Following Gaussian linear transformation properties, the resulting marginal density is Gaussian with known parameters, provided that the likelihood mean, Qz_n, is represented as a linear transformation over z_{kn}. Using ordinary matrix algebra, we have decomposed the mean Qz_n into two terms, $Qz_n = q_k z_{kn} + [Qz_n]_{z_{kn}=0}$. The former includes the parameter of interest z_{kn}, whereas the latter (residue term) excludes z_{kn} through a reduced product $[Qz_n]_{z_{kn}=0}$ 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 6. As can be seen in N_l, the conditional probability $P(y_n|x_n)$ models the correlations among response variables (y_{1n} to $y_{D_q n}$) thanks to the full covariance matrix, $\Psi_y + q_k {\sigma_{zk}}^{-1} q_k^T$. 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_l/D_l ratio.

$$
\begin{aligned}
N_l &= \int \mathcal{N}(y_n|Qz_n, \Psi_y)\mathcal{N}(z_{kn}|p_k x_n, \sigma_{zk}) dz_{kn} \\
&= \mathcal{N}(y_n|q_k p_k x_n + [Qz_n]_{z_{kn}=0}, \Psi_y + q_k {\sigma_{zk}}^{-1} q_k^T)
\end{aligned} \tag{6}
$$

$$
D_l = \mathcal{N}(y_n|[Qz_n]_{z_{kn}=0}, \Psi_y)
$$

The prior ratio (r_p) is derived below, following Equation 2. 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. 6 and 7), the posterior ratio for existing features may be derived as a uniform draw over $\frac{r}{r+1}$.

$$
r_p = \frac{m_{k,-i}}{N - 1 - m_{k,-i}} \tag{7}
$$

Next, we should decide the number of new features (κ_n) added for the current observation. IBP implies that κ_n is *a priori* distributed as $Poisson(\frac{\alpha}{n})$. Yet, there have been different approaches for sampling κ_n from its posterior. In (Griffiths and Ghahramani, 2011) κ_n is sampled through a MAP estimate of various values for κ_n, ranging from zero to an upper bound. However, (Knowles and Ghahramani, 2011) proposes a Metropolis-Hastings step to some random κ_n and evaluates the acceptance through posterior ratios. Each MH jump from $\xi \to \xi^*$ is done with a probability $J(\xi^*|\xi)$ with varying underlying assumptions. A basic approach considers the prior on ξ^* as candidate function, i.e. $J(\xi^*|\xi) = P(\xi)P(\xi^*)$.

$$
\begin{aligned}
r_{\xi \to \xi^*} &= \frac{P(\xi^*|y_n, \_)P(\xi^*)P(\xi)}{P(\xi|y_n, \_)P(\xi)P(\xi^*)} \\
&= \frac{\mathcal{N}(y_n|q'_k p'_k x_n + [Q^* z_n]_{z'_{kn}=0}, \Psi_y + q'_k {\sigma_{zk}}^{-1} {q'_k}^T)}{\mathcal{N}(y_n|[Qz_n]_{z'_{kn}=0}, \Psi_y)}
\end{aligned} \tag{8}
$$

In the equations above, _ denotes the remaining parameters and * superscripts indicate extended parameters with κ_n new features. The jump is ultimately accepted with probability $min(1, r_{\xi \to \xi^*})$. In order to add κ_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 κ_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. 1), 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, $P(q_k|Y, Q^{-k}, Z, \Psi_y, \Psi_q) = \prod_{i=1}^{D_q} \mathcal{N}(q_{ik}|\mu_{ik}, \Sigma_{ik}), \forall q_{ik} : i \in [1, D_q]$. Following a similar approach, we derive the posterior for p_k.

$$
\begin{aligned}
&\mathcal{N}(q_{ik}|\mu_{q_{ik}}, \sigma_{q_{ik}}) \propto \\
&\mathcal{N}(y_{i:}|q_{ik} z_k + [q_{i:} Z]_{q_{ik}=0}, \sigma_{yi})\mathcal{N}(q_{ik}|0, \sigma_{qi}) : \\
&\sigma_{q_{ik}} = \frac{\sigma_{qi}\sigma_{yi}}{\sigma_{yi} + \sigma_{qi} z_k z_k^T}, \mu_{q_{ik}} = \frac{\sigma_{qi}(y_{i:} - [q_{i:} Z]_{q_{ik}=0}) z_k^T}{\sigma_{yi} + \sigma_{qi} z_k z_k^T} \\
\\
&\mathcal{N}(p_k|\mu_{p_k}, \Sigma_{p_k}) \propto \mathcal{N}(z_k|p_k X, \sigma_{zk})\mathcal{N}(p_k|0, \sigma_{pj} I_{D_p}) : \\
&\Sigma_{p_k} = \sigma_{pj} I_{D_p} + \sigma_{zk}(XX^T)^{-1}, \mu_{p_k}^T = \Sigma_{p_k} \frac{X z_k^T}{\sigma_{zk}}
\end{aligned} \tag{9}
$$

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.

$$\begin{aligned}&N(z_{kn}|\mu_{z_{kn}}, \sigma_{z_{kn}}) \propto \\ &\mathcal{N}(y_n|q_k z_{kn} + [Qz_n]_{z_{kn}=0}, \Psi_y)\mathcal{N}(z_{kn}|p_k x_n, \sigma_{zk}) : \\ \\ &\sigma_{z_{kn}} = (\frac{1}{\sigma_{zk}} + q_k^T \Psi_y^{-1} q_k)^{-1}, \\ &\mu_{z_{kn}} = \sigma_{z_{kn}}[q_k^T \Psi_y^{-1}(y_n - [Qz_n]_{z_{kn}=0}) + \frac{p_k x_n}{\sigma_{zk}}]\end{aligned} \tag{10}$$

4.3 Noise and factor loading covariances and IBP parameters

NCFR adopts diagonal noise models, Ψ_y and Ψ_z. The factor loading covariances, Ψ_q and Ψ_p, are also diagonal. Following (Knowles and Ghahramani, 2011), 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 α is sampled through a conjugate Gamma(e, f) prior, as follows. The distribution for $P(S|\alpha)$ and other notations are introduced in section 2.1.

$$\begin{aligned}&IG(\sigma_{yi}|a + \frac{N}{2}, b + tr(E_{yi}^T E_{yi})) \propto \mathcal{N}(E_{yi}|0, \sigma_{yi})IG(\sigma_{yi}|a, b) \\ &IG(\sigma_{zk}|a + \frac{m_k}{2}, b + tr(E_{zk}^T E_{zk})) \propto \mathcal{N}(E_{zk}|0, \sigma_{zk})IG(\sigma_{zk}|a, b) \\ &IG(\sigma_{qi}|c + \frac{D_q}{2}, d + tr(q_k q_k^T)) \propto \mathcal{N}(q_k|0, \sigma_{qi})IG(\sigma_{qi}|c, d) \\ &IG(\sigma_{pj}|c + \frac{D_p}{2}, d + tr(p_k^T p_k)) \propto \mathcal{N}(p_k|0, \sigma_{pj})IG(\sigma_{pj}|c, d) \\ &G(\alpha|K + e, f + H_N) \propto Poisson(S|\alpha)G(\alpha|e, f)\end{aligned} \tag{11}$$

5 Experiments

Following the blueprint set forth in sections 1 and 4.1, we compare the models listed in Table 1. Each column represents a model category and its related variants. The abbreviations in italics are used for easier referencing. For comparison, we have implemented a baseline EM solution for conditional factor regression (EM-CFR), with unit noise and loading variances (for details, refer to the supplementary material). This solution is indicated as a baseline, along with a slightly more complex MCMC implementation with diagonal noise and loading covariances. Since the number of dimensions in the low-rank latent layer needs to be predefined, we have tried them with three different Ks, according to the optimal value for each experiment. Under the third column, the variants of our proposed model (NCFR) are listed, pivoted arount the *generic* solution as specified in the model section. Thanks to the IBP, we have initialised K_0 in the first iteration with a random IBP-generated value instead of having to choose an arbitrary number. All experiments are run over 10 trials with $\alpha_0 = 0.01$ The code for NCFR and related models is inspired by (Knowles, Accessed 2013) and consists of configuration tools to run most of the following models, soon to be available online.

Algorithm 1: NCFR posterior sampling algorithm

```
Input: $\alpha_0, X, Y$
Output: $P(\Phi|Y, X)$ :
    $\Phi = \{Q, P, S, Z, \Psi_y, \Psi_z, \Psi_q, \Psi_p, \alpha\}$,
    $\Phi_{MAP}$ for prediction
Initialize: All variables by their priors,
including $S$ with $K_0$ factors derived by IBP($\alpha$)
for iteration = 1 to ConvergenceMax do
    forall the $s_{k,n} \in S$ do
        // sample $s_{k,n}$, if active also sample $z_{k,n}$
        Sample $s$ and $z$ (k,n)      // Equations 5-8,10
    end
    forall the $q_{:k} \in Q$ do
        Sample $q_k$ (k)              // Equation 9
    end
    forall the $p_{k:} \in P$ do
        Sample $p_k$ (k)              // Equation 9
    end
    Sample $\Psi_y, \Psi_z, \Psi_q, \Psi_p, \alpha$    // Equation 11
end
return $P(\Phi|Y, X)$ and $\Phi_{MAP}$ for prediction
```

Similarly to projection/factor loading matrices in PPCA and factor analysis (Tipping and Bishop, 1999), the loading matrices of our proposed model are not fully identifiable. Unidentifiability is avoided in some studies by constraining the loading matrices to be lower triangular (West, 2003). However, such treatments are not necessary for prediction purposes (Bhattacharya and Dunson, 2011) and usually used for parameter comparison. Instead, we have utilised prediction likelihood and normalised prediction error for performance measurement. Overfitting is monitored by auditing K, mixing efficiency is examined for the MCMC sampler and a final study on computational speed proves the viability of our solution.

5.1 Synthetic data

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 Tenorth et al., 2009). Z is calculated

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 *(EMCFRl)*	Generic *(Gen)*
	EM, medium K *(EMCFRm)*	Fixed α *(Fixα)*
	EM, small K *(EMCFRs)*	Isotropic noise *(IsoN)*
	MCMC, large K *(CFRl)*	Isotropic loading *(IsoL)*
	MCMC, medium K *(CFRm)*	
	MCMC, small K *(CFRs)*	

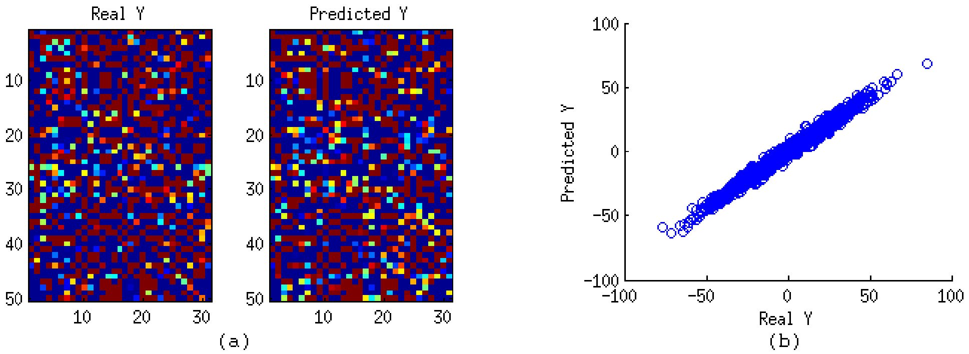

Figure 2: *IsoN* NCFR performance on synthetic data: (a) Heat map for visual comparison of real $Y_{(D_q \times N_{test})}$ matrix vs. predicted Y. (b) Scatter plot of real Y vs. predicted Y with all dimensions superimposed on the same plot. The plot includes $D_q \times N_{test}$ points all aligned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

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 Ψ_y (eq. 3). 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 ($\epsilon = \| y_n - \tilde{y}_n \| / \| y_n \|$) are reported in Table 2, 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 the accuracy in *IsoN* variant via heat maps and scatter plots of predicted Y vs. real Y (fig. 2). The scatter plot also shows significant accuracy, aligning the points along the line $y = x$ (fig. 2(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 (Tenorth et al., 2009), 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 oc-

Table 2: Prediction error percentage ($\epsilon = 100\times \parallel y_n - \tilde{y}_n \parallel / \parallel y_n \parallel$) for synthetic prediction.

Classic Linear Regression		Conditional Factor Regression		Non-parametric CFR	
FRR	370.52 ± 259.34	*CFRl*	6.68 ± 2.90	***Gen***	**6.50 ± 2.83**
		CFRm	6.64 ± 2.90	*Fixα*	88.55 ± 18.32
		CFRs	89.23 ± 14.57	***IsoN***	**6.60 ± 2.76**
				IsoL	6.73 ± 2.88

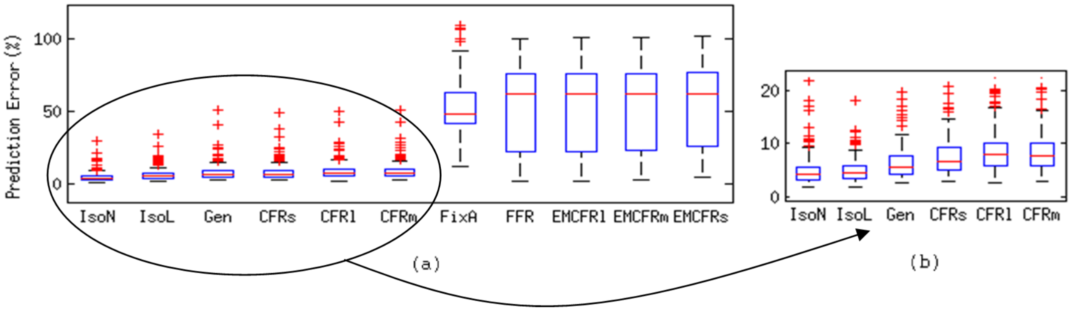

Figure 3: Boxplot of prediction performance on TUM kitchen data, sequence 1-2: (a) Normalised prediction errors are presented over 148 frames of Y and sorted across the 3 categories of models. The top 6 variants are magnified in the plot on the right (b), showing noticeably better performance for NCFR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

curences of the action "Carrying while walking" and tested it over an unseen occurence, using sequences performed by Subject1: *1-1* to *1-5*.

Figure 3(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, making pose prediction more challenging for those frames. This explains the few outlier errors in Figure 3.

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 (EIA) 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 $X_{(25\times24)}$ and $Y_{(25\times24)}$ 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 (West, 2003). The forecast results are reported in Table 3, 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 5.4. 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%.

Table 3: Prediction error percentage ($\epsilon = 100\times \parallel y_n - \tilde{y}_n \parallel / \parallel y_n \parallel$) for gas consumption forecast.

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

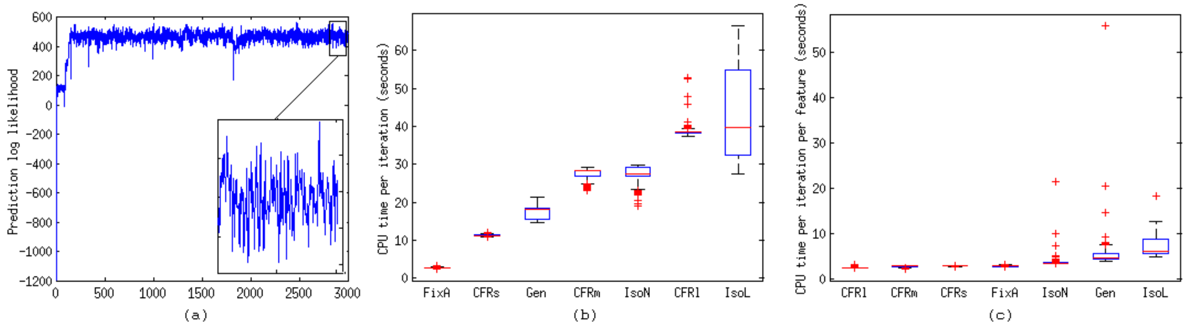

Figure 4: Sampling efficiency and computational cost: (a) Log likelihood plot of NCFR gas results with the last 200 iterations magnified, showing convergence and well mixing. (b) Inference time per iteration for TUM kitchen pose prediction (c) Inference time per iteration per active feature for the same 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.

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 4(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 viability. Figures 4(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 resiliance and adaptability against noise and overfitting.

References

Energy information administration (eia) of the united states: Natural gas monthly archive. http://www.eia.gov/naturalgas/monthly/?src=Natural-f3. Accessed: 2013-09-20.

A. Bhattacharya and D. B. Dunson. Sparse bayesian infinite factor models. *Biometrika*, 98(2):291–306, 2011. doi: 10.1093/biomet/asr013. URL http://biomet.oxfordjournals.org/content/98/2/291.abstract.

L. Bo and C. Sminchisescu. Supervised spectral latent variable models. In *International Conference on Artificial Intelligence and Statistics*, volume 246, page 248, 2009.

F. Caron and A. Doucet. Sparse bayesian nonparametric regression. In *Proceedings of the 25th international conference on Machine learning*, pages 88–95. ACM, 2008.

C. M. Carvalho, J. Chang, J. E. Lucas, J. R. Nevins, Q. Wang, and M. West. High-dimensional sparse factor modeling: Applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484):1438–1456, 2008. doi: 10.1198/016214508000000869. URL http://amstat.tandfonline.com/doi/abs/10.1198/016214508000000869.

T. Griffiths and Z. Ghahramani. Infinite latent feature models and the indian buffet process. In Y. Weiss, B. Schölkopf, and J. Platt, editors, *Advances in Neural Information Processing Systems 18*, pages 475–482. MIT Press, Cambridge, MA, 2006.

T. L. Griffiths and Z. Ghahramani. The Indian buffet process: An introduction and review. *J. Mach. Learn. Res.*, 999999:1185–1224, July 2011. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=2021026.2021039.

D. Knowles. Infinite sparse factor regression. http://mlg.eng.cam.ac.uk/dave/isfa.zip, Accessed 2013. Accessed: 2013-05-31.

D. Knowles and Z. Ghahramani. Infinite sparse factor analysis and infinite independent components analysis. In *Proceedings of the 7th international conference on Independent component analysis and signal separation*, ICA'07, pages 381–388, Berlin, Heidelberg, 2007. Springer-Verlag. ISBN 3-540-74493-2, 978-3-540-74493-1. URL http://dl.acm.org/citation.cfm?id=1776684.1776735.

D. Knowles and Z. Ghahramani. Nonparametric Bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, 5(2B):1534–1552, 2011.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

W. Kusakunniran, Q. Wu, J. Zhang, and H. Li. Support vector regression for multi-view gait recognition based on local motion feature selection. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 974–981, 2010. doi: 10.1109/CVPR.2010.5540113.

S. Montagna, S. T. Tokdar, B. Neelon, and D. B. Dunson. Bayesian latent factor regression for functional and longitudinal data. *Biometrics*, 68(4):1064–1073, 2012. ISSN 1541-0420. doi: 10.1111/j.1541-0420.2012.01788.x. URL http://dx.doi.org/10.1111/j.1541-0420.2012.01788.x.

P. Rai and H. Daumé III. Multi-label prediction via sparse infinite cca. *Advances in Neural Information Processing Systems*, 22:1518–1526, 2009.

Y. W. Teh, M. Seeger, and M. I. Jordan. Semiparametric latent factor models. In R. G. Cowell and Z. Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, Jan 6-8, 2005, Savannah Hotel, Barbados*, pages 333–340. Society for Artificial Intelligence and Statistics, 2005. (Available electronically at http://www.gatsby.ucl.ac.uk/aistats/).

M. Tenorth, J. Bandouch, and M. Beetz. The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition. In *IEEE International Workshop on Tracking Humans for the Evaluation of their Motion in Image Sequences (THEMIS), in conjunction with ICCV2009*, 2009.

M. E. Tipping and C. M. Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 61:611–622, 1999.

Y. Wang and J. Zou. Vast volatility matrix estimation for high-frequency financial data. *The Annals of Statistics*, 38(2):943–978, 2010.

M. West. Bayesian factor regression models in the large p, small n paradigm. *Bayesian statistics*, 7(2003): 723–732, 2003.