Latent feature regression for multivariate count data

Arto Klami[1] **Abhishek Tripathi**[2] **Johannes Sirola**[1] **Lauri Väre**[1] **Frederic Roulland**[3]
[1]Department of Computer Science, Helsinki Institute for Information Technology HIIT, University of Helsinki
[2]Xerox Research Centre India [3]Xerox Research Centre Europe

Abstract

We consider the problem of regression on multivariate count data and present a Gibbs sampler for a latent feature regression model suitable for both under- and overdispersed response variables. The model learns count-valued latent features conditional on arbitrary covariates, modeling them as negative binomial variables, and maps them into the dependent count-valued observations using a Dirichlet-multinomial distribution. From another viewpoint, the model can be seen as a generalization of a specific topic model for scenarios where we are interested in generating the actual counts of observations and not just their relative frequencies and co-occurrences. The model is demonstrated on a smart traffic application where the task is to predict public transportation volume for unknown locations based on a characterization of the close-by services and venues.

1 INTRODUCTION

Multivariate regression refers to the problem of learning a regression model from D-dimensional covariates $\mathbf{x}$ to L response variables $\mathbf{y} = [y_1, ..., y_L]$. Correlated real-valued responses can be modeled with multivariate normal and t-distributed noise, but count responses will need L separate models since there is no efficient closed-form multivariate distribution over counts. Even if the learning tasks are tied together by regularizing the regression weights of the L different learners in a multi-task learning fashion, the predictions are still independent over the dimensions.

An alternative approach to learning multivariate regression models is via latent representation, which correlates also the predictions. By first learning a mapping from $\mathbf{x}$ to a set of K latent features z_k and then from those to the L observed features, we can construct multivariate predictive distributions without needing closed-form multivariate distributions. This kind of a strategy is widely used in reduced-rank regression [Izenman, 1975], where the latent features are of lower dimensionality than the inputs or the outputs, and it also generalizes beyond Gaussian models.

There are, however, only a few multivariate regression solutions specifically for count data. For univariate count data several solutions have been presented, but for multivariate responses there are few dedicated tools besides vector generalized linear models [Yee and Wild, 1996] and their reduced-rank extensions [Yee and Hastie, 2003]. The main goal of our work is to provide a practical Bayesian regression tool to fill this vacuum, building on the recent advances in Bayesian inference for negative binomial distributions [Polson et al., 2013, Zhou et al., 2012].

Our model solves the regression problem by introducing K latent features following negative binomial (NB) distribution. These latent features are considered as response variables for a regression layer and as input variables for another layer mapping the latent features into the observed counts using a Dirichlet-multinomial distribution. Intuitively, the model can be thought of as generating counts for unknown processes, which in turn distribute the counts across the observed variables. While $K \ll \min(D, L)$ is a natural choice for this kind of models, we do not pose the solution as that of reduced-rank regression since we also considers models with large K. This makes the model family flexible enough to describe both overdispersed data (variance larger than the mean) and underdispersed data (variance smaller than the mean).

The proposed latent feature regression model has another interesting interpretation: It is a topic model that accepts arbitrary covariates for the topic probabilities similar to the Dirichlet-multinomial regression (DMR) model by Mimno and McCallum [2008],

Appearing in Proceedings of the 18^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2015, San Diego, CA, USA. JMLR: W&CP volume 38.

but that generates the actual counts of words instead of just their relative frequencies. In a typical setting, topic models describe a set of text documents as a collection of topics that are distributions over the word tokens, providing an interpretable summary of the data. However, they generate the topic counts conditional on the total document length, which corresponds to assuming that the topic distribution is independent of the length. While this is a reasonable assumption in many topic modeling applications, it does not strictly hold even for text analysis. For example, Doyle and Elkan [2009] discussed topic models that take into account word burstiness, the observation that words already appearing in a document are more likely to occur again, which implies that the choices are not independent. Our model, when interpreted as a topic model, is a direct generalization of DMR without this simplifying assumption. Instead, it explicitly generates the counts of the word tokens and can hence be applied to arbitrary count regression problems.

We apply the proposed model to data that combines both over- and underdispersion: We model public transportation (bus) volume conditional on covariates describing the neighborhood of a bus stop, asking the question of how well the number of passengers boarding a bus at a given time can be predicted solely based on static demographic information and training data collected from other stops. For high-volume stops the data is underdispersed, whereas for low-volume stops it is overdispersed. The model outperforms both vector generalized linear models and independently computed regression models for the L dimensions.

2 BACKGROUND

2.1 Latent Feature Regression

Reduced-rank regression refers to models that solve the multivariate regression problem from $\mathbf{x} \in \mathbb{R}^D$ to $\mathbf{y} \in \mathbb{R}^L$ as $\mathbf{y} \approx \mathbf{W}\mathbf{x}$ by learning the weight matrix $\mathbf{W} \in \mathbb{R}^{L \times D}$ as a low-rank product of two matrices: $\mathbf{W} = \mathbf{U}\mathbf{V}$, where $\mathbf{U} \in \mathbb{R}^{L \times K}$ and $\mathbf{V} \in \mathbb{R}^{K \times D}$. Compared to directly formulating $\mathbf{W}$ as a $L \times D$ matrix this representation typically has fewer parameters and better accuracy [Izenman, 1975].

Such a model can also be interpreted as a two-stage regression that has K-dimensional latent representation for the samples. These latent variables are computed with one regression model as $\mathbf{z} \approx \mathbf{V}\mathbf{x}$, and the outputs are then modeled with another regression layer as $\mathbf{y} \approx \mathbf{U}\mathbf{z}$. We present our regression model using this formulation, explicitly representing the latent features, not only because it naturally fits the generative modeling approach but also because we do not limit to low-rank models but also consider over-complete solutions where K can be larger than L and/or D. In Section 4 we will show why such models can be useful when modeling underdispersed count data.

2.2 Negative Binomial Models

We model the counts with the negative binomial distribution

$$\mathrm{NB}(y|r,p) = \frac{\Gamma(y+r)}{y!\Gamma(r)}(1-p)^r p^y, \tag{1}$$

which can alternatively be expressed as a mixture of Poisson distributions using a gamma prior with shape r and scale $p/(1-p)$ on the Poisson rate parameter. The advantage of NB compared to Poisson is that we can tune the mean and variance of the outcome using r and p, instead of assuming them to be equal. However, the variance cannot be made smaller than the mean to model underdispersed data.

Efficient inference for NB models has previously been challenging, but two recent results have made NB-based count models practical. Polson et al. [2013] presented an auxiliary variable augmentation strategy that re-writes the likelihood (1) as a mixture over Pólya-Gamma variables. By parameterizing $p = \mathrm{logistic}(\psi) = (1+e^{-\psi})^{-1}$ we get the mixture density

$$\begin{aligned}\mathrm{NB}(y|r,p) &\propto \frac{(e^{\psi})^y}{(1+e^{\psi})^{y+r}} \\ &= \frac{e^{(y-r)/2}}{2^{y+r}} \int e^{-\omega\psi^2/2}\mathrm{PG}(\omega|y+r,0)d\omega,\end{aligned}$$

where $\mathrm{PG}(\omega|b,c)$ is the Pólya-Gamma distribution. With Gaussian priors on ψ, explicitly instantiating ω leads to an auxiliary variable sampler with closed-form Gibbs updates for both ψ and ω.

The other crucial result is by Zhou and Carin [2012], who derived another augmentation strategy for inferring the parameter r using a compound-Poisson representation for the NB distribution. They use an auxiliary variable l to record the number of tables occupied in a Chinese restaurant process with y customers and concentration parameter r. Given a gamma prior on r, the conditional distribution $r|y,l$ is also gamma.

Combining these two results provides a state-of-the-art univariate count regression model [Zhou et al., 2012], denoted by LGNB for lognormal and gamma mixed negative binomial regression, which we will use as one of the baselines when evaluating the proposed model.

2.3 Topic models

Topic models are based on the simple generative process, where every item i of an object d chooses topic z_i

and then a token w_i given the topic [Blei et al., 2003]. Often the objects are called documents, the tokens are possible words of a language, and the items are the individual words in the document. In this work, however, we use the more general terminology since the main application of our model is not in modeling text.

The most common choices for the distributions are given by the Dirichlet-multinomial

$$z_i \sim \text{Mult}(\boldsymbol{\theta}_d), \qquad w_i|z_i = k \sim \text{Mult}(\boldsymbol{\phi}_k),$$
$$\boldsymbol{\theta}_d \sim \text{Dir}(\zeta), \qquad \boldsymbol{\phi}_k \sim \text{Dir}(\gamma).$$

Here $\boldsymbol{\theta}_d$ is a distribution over the topics for item d and $\boldsymbol{\phi}_k$ is a distribution over the tokens for topic k. The model is implicitly conditioned on the total number of items m_d for each object. The early papers briefly discussed this issue and for example Blei et al. [2003] explicitly modeled the total length as $m_d \sim \text{Poisson}(\lambda)$. Since this count is independent of the remaining parameters, effectively all topic models nowadays ignore the total length.

Gibbs sampling for topic models is typically performed by marginalizing out the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$, enabling direct sampling of individual token-to-topic assignments conditional on all other assignments:

$$p(z_{di} = k|-) \propto (\zeta + m_{dk.})\frac{m_{.kw} + \gamma}{W\gamma + m_{.k.}}.$$

Here m_{dkw} is the number of tokens w in item d assigned to topic k, excluding the current sample, dots denote summation over the corresponding index, and W is the vocabulary size. The sampler iteratively re-samples z_{di} and updates the counts accordingly.

The Dirichlet multinomial regression model by Mimno and McCallum [2008] replaces the independent topic probabilities $\boldsymbol{\theta}_d \sim \text{Dir}(\zeta)$ with probabilities that are conditional on arbitrary set of features $\mathbf{x}_d$:

$$\boldsymbol{\theta}_d \sim \text{Dir}(\boldsymbol{\zeta}_d),$$
$$\zeta_{dk} = \exp(\boldsymbol{\beta}_k^T \mathbf{x}_d).$$

The parameters of the (now non-symmetric) Dirichlet distribution are modeled conditional on the features, and a normal prior is given for the regression weights $\boldsymbol{\beta}_k$. For inference they used generic numeric optimization to find the maximum a posteriori estimate of the regression weights while updating the rest of the model parameters with Gibbs sampling. The numerical optimization requires computing the gradients with respect to $\boldsymbol{\beta}$ which involves somewhat heavy evaluation of digamma-functions.

Conditioning the topic proportions on covariates allows discovering topics that correspond to, for example, specific authors or sources by providing the author and source information as covariates [Mimno and McCallum, 2008]. Yuan et al. [2012] used the model for an application resembling ours, to understand (but not to predict) mobility patterns in a city.

3 MODEL

3.1 Generative process

We build a latent feature regression model for count data by combining the elements presented above. We provide for each sample a latent vector $\mathbf{z} \in \mathbb{N}^K$ consisting of non-negative counts. These counts are modeled with negative-binomial regression from the covariates $\mathbf{x}$. The actual observations $\mathbf{y}$ are constructed by distributing the latent counts over the L observed features by Dirichlet-multinomials. This formulation not only naturally models overdispersed counts but also underdispersed counts; the observed count for y_l is obtained by summing over the subset of the K latent topics that have non-negligible probability for generating tokens for that dimension, and this summation naturally decreases the variance by a factor of the subset size.

The model, illustrated in Figure 1, is specified as

$$\begin{aligned}
p_{nk} &= \frac{e^{\psi_{nk}}}{1+e^{\psi_{nk}}}, \qquad \psi_{nk} = \text{logit}(p_{nk}) = \boldsymbol{\beta}_k^T \mathbf{x}_n + \epsilon_{nk}, \\
\epsilon_{nk} &\sim \text{N}(0, \tau_k^{-1}), \qquad \tau_k \sim \text{Gamma}(f_0, 1/g_0), \\
z_{nk} &\sim \text{NB}(r_k, p_{nk}), \qquad y_{nj}|t_{nj} = k \sim \text{Mult}(\boldsymbol{\phi}_k), \\
\boldsymbol{\beta}_k &\sim \prod_{d=1}^{D} \text{N}(0, \alpha_{dk}^{-1}), \qquad \alpha_{dk} \sim \text{Gamma}(c_0, 1/d_0), \\
r_k &\sim \text{Gamma}(a_0, 1/h_k), \qquad h_k \sim \text{Gamma}(b_0, 1/e_0), \\
\boldsymbol{\phi}_k &\sim \text{Dir}(\gamma),
\end{aligned} \tag{2}$$

where $a_0, b_0, c_0, d_0, e_0, f_0$ and g_0 are constant hyperparameters which we set to small values, n denotes the sample and j indexes the items contained in each $\mathbf{y}_n$. The additive Gaussian noise term ϵ_{nk} can be interpreted as lognormal noise for the NB distribution.

The crucial elements are the K NB regression models that generate the latent counts z_{nk} from the covariates, parameterized by the regression weights $\boldsymbol{\beta}_k$ and rates r_k, and the K topic distributions that distribute the counts over the observed dimensions y_l, parameterized by $\boldsymbol{\phi}_k$. The auxiliary vectors $\mathbf{t}_n \in [1, \ldots, K]^{y_n}$ are introduced solely for inference purposes; these are indicators telling the topic assignments of individual tokens and the counts z_{nk} can be deterministically constructed as $z_{nk} = \sum I[t_{nj} = k]$.

3.2 Inference

For inference we perform partially collapsed Gibbs sampling, utilizing the auxiliary variable constructions

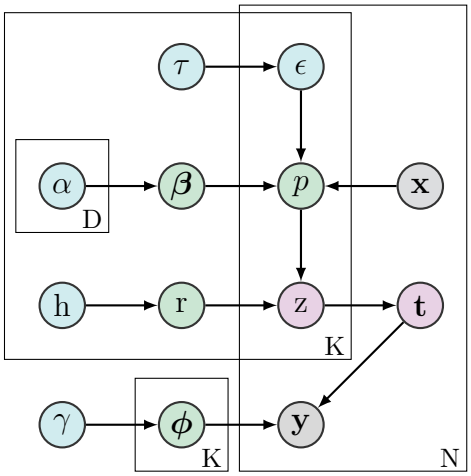

Figure 1: Plate diagram of latent feature regression. The covariates $\mathbf{x}$ are mapped into latent counts $\mathbf{z}_k$ via negative binomial regression using logistic transformation for $p_k = \text{logistic}(\boldsymbol{\beta}_k^T \mathbf{x} + \epsilon_k)$, and the counts are then distributed across the L output dimensions to create $\mathbf{y}$ via token-assignment vectors $\mathbf{t}$.

by Zhou and Carin [2012] for inferring $\boldsymbol{\beta}_k$ and r_k. Following Klami [2014], we also consider a model variant that omits the lognormal noise in (2) by setting $\tau_k = \infty$; this improves mixing since $\boldsymbol{\beta}_k$ now directly depends on $\mathbf{z}$, but results in slightly slower updates and different predictive distribution. The Gibbs updates for both variants are provided in the Appendix.

The inference for z_{nk} is done by collapsing $\boldsymbol{\phi}_k$, but conditional on r_k and p_{nk}. The variables z_{nk} are sampled implicitly by re-sampling the token assignments t_{nj} of the current sample. We exclude one sample at a time and pick a new token from a multinomial

$$p(t_{nj} = k | y_{nj} = w, -) \propto \frac{r_k + z_{nk}}{1 + z_{nk}} p_{nk} \cdot \frac{\gamma + m_{kw}}{L\gamma + m_{k.}},$$

where m_{kw} is the number of tokens assigned to output dimension w generated by topic k. The first part of this term, including p_{nk}, comes from arithmetic simplification of ratio of NB densities, whereas the latter corresponds to the Dirichlet-multinomial.

The computational complexity of the sampler is linear as a function of K and N, and cubic as a function of D due to inversion of the covariance of $\boldsymbol{\beta}_k$. Overall, the computational demand is roughly comparable to learning K independent LGNB regression models [Zhou et al., 2012]. For multivariate setups the model is hence faster than LGNB when learning a reduced-rank solution, especially for very high-dimensional output spaces; our model has a running time independent of L, whereas naive comparison methods need to be run for each output dimensions separately. The sampling of the token assignments, however, makes the algorithm scale linearly in the total number of tokens; in our experiments this was not a computational bottleneck, but for extremely large counts the updates could be done directly for $\mathbf{z}$ to improve speed.

3.3 Related Work

From the modeling perspective the work is most closely related with the LGNB regression model by Zhou et al. [2012], extending it to multivariate cases. We use the core elements of LGNB for the first stage of the model, for regressing from the covariates to the latent variables, but additionally present inference details also for a related model that omits the lognormal noise element. Besides extending LGNB for multivariate regression, the proposed model is more accurate also for univariate regression problems when the outputs are underdispersed, as will be shown in Section 4.

The existing multivariate regression models for count data fall largely into the family of vector generalized linear models (VGLM) [Yee and Wild, 1996], and their reduced-rank extensions [Yee and Hastie, 2003]. VGLMs are a very flexible family of tools, but based on maximum likelihood estimation and lack robustness compared to our Bayesian analysis. The same limitation applies to the probabilistic multivariate count models studied by Ghitany et al. [2012]; they learn maximum likelihood estimate with expectation maximization. Ma et al. [2008] proposed also Bayesian multivariate models, but only for two to three dimensions and using inefficient generic MCMC sampling. In the experiments we show how our model outperforms non-regularized VGLM as well as regularized standard GLMs applied for each output variable at a time; we are not aware of regularized VGLMs.

The model is also an extension of Dirichlet-multinomial regression (DMR) [Mimno and McCallum, 2008]. The crucial difference is that our model generates the actual observed counts instead of just the relative frequencies, making it applicable for problems where the counts themselves are required. To solve a regression problem, DMR needs to be complemented with a separate model for the total count, similar to how standard topic models need a (non-informative) model $n \sim \text{Poisson}(\lambda)$ to generate the total length. In the experimental section we demonstrate such an approach by combining their model with LGNB regression for the total count and show that it does not result in accurate predictions. This is, at least in part, because of the gross simplifying assumption that the total length is independent of the topic distribution, but the gradient-based maximum a posterior estimation for $\boldsymbol{\beta}_k$ could also play a role. On the other hand, DMR seems to typically result in sparser latent count

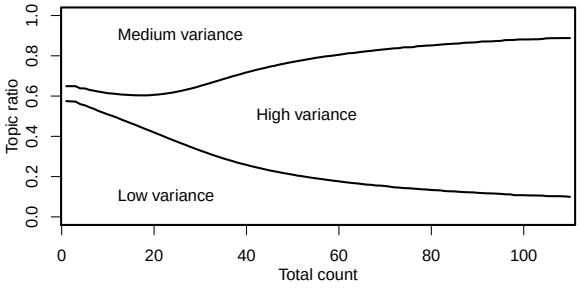

Figure 2: The model allows the topic ratios to depend on the total length of the document, so that low-variance topics generate topics for almost all documents but never create very many of them, whereas high-variance topics either do not generate any tokens or generate several at once. Consequently, the latter types dominate documents that are long. An example of a low-variance topic would be one that generates structural words ("keywords", "abstract", "citations") for scientific documents, whereas controversial topics that can create length discussions could show up as high-variance topics in analysis of online news comments. For a regular topic model the ratios would be constant lines, since they assume the topic probability to be independent of the document length.

distributions compared to our model, which might be beneficial in some applications.

4 MODEL PROPERTIES

Next we illustrate two basic properties of our model. The first illustration is related to the interpretation of the model as a topic model conditional on arbitrary covariates. It demonstrates the difference between explicitly generating the counts of individual topics instead of generating the total count of tokens (which is usually left implicit in the topic modeling literature) and then distributing the tokens into the topics according to fixed topic proportions. Figure 2 shows how the relative ratio for three topics with the same mean (each generates on average 10 tokens) but different variance behaves as a function of the total count. The high-variance topic dominates for large total counts and the low-variance one for small ones, as expected. For regular topic models the ratios would be constants of $1/3$ irrespective of the total count. While that assumption can hold in some applications, it is still a simplification in the general case.

The second illustration shows how we can flexibly model both under- and overdispersed count data. While majority of the literature on modeling count data is concentrated on proper treatment of overdispersed data, several practical data sources show underdispersion. This is the case especially when modeling systems where an external agent balances the counts by prior design or by actively reacting to the counts, for example by opening new counters to reduce the number of clients lining up for service. As another example, the number of cars driving along a highway during a rush hour is underdispersed; the count is high but the variance is small because the road cannot accommodate more cars than its soft capacity.

We illustrate the model on two simple artificial data sets. Each of 100 data points is assigned to one of 4 different patterns so that the covariates for each pattern are given as noise-corrupted one-hot codes and each pattern generates observations for one output dimension. This implies that the output dimensions are actually independent. We present results for two data sets, one generating underdispersed counts using the Conway-Maxwell-Poisson distribution, and the other generating overdispersed data from NB distribution. Figure 3 compares the proposed model in this task to the alternative of learning 4 separate LGNB models, one for each output dimension. For overdispersed data LGNB is as accurate as our model, as is to be expected; the problem consists of learning L independent models and the data generating process matches the models. For underdispersed data our model with large K outperforms LGNB by distributing the generation process across multiple topics.

5 MODELING PUBLIC TRANSPORTATION VOLUME

Understanding and optimizing peoples mobility has become a concern for cities authorities: the growing traffic congestion and pollution has a significant impact on the daily productivity and perceived quality of life of citizens. Nowadays intelligent transportation systems collect vast amounts of usage data, and they routinely predict transportation demand at existing stops and routes. However, such models are not helpful for estimating passenger counts on potential new routes and/or stops, since they are ultimately based on modeling the current mobility patterns.

In order to accurately understand the public transportation demand, we need not only to know what are the current flows but also what are the reasons for these flows. Understanding activities that generate a demand for mobility can help build better models to predict demand at potential new routes and/or stops; for example Gutierrez et al. [2011] predicted metro transit counts from demographic properties of the neighborhoods. Here, we consider service and venue characteristics of the area surrounding the stop and other stops that can be reached from this stop to

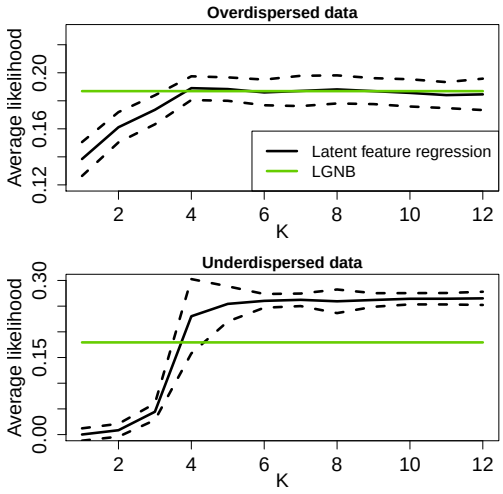

Figure 3: Regression accuracy, measured by mean predictive likelihood averaged over 50 data sets for two types of data. For overdispersed data (top) with L=4 independent dependent variables the latent feature regression model reaches the accuracy of independent LGNB models as soon as it has enough latent features, as it should. For underdispersed data (bottom) it outperforms LGNB already for $K = 4$, and the overcomplete solutions with $K > L$ provide still higher accuracy by reducing the variance of the predictions to better match the data variance. The dashed lines indicate one standard deviation in LFR accuracy relative to LGNB.

understand it's demand and to estimate roughly how much information is contained in this partial view.

5.1 Data

We collected data from an undisclosed city during 21 days. We present the traffic volumes along 37 different routes and 603 stops as 51-dimensional vectors $\mathbf{y}$ that describe the number of passengers who boarded a bus on that route at that stop during each 20-minute interval between 5am and 10pm (the passenger volume outside this window is negligible). We study the predictions at granularity of 20 minutes to make the comparison methods feasible; they would become computationally too heavy for smaller windows whereas our method is independent of the window size and could just as well make predictions on a minute scale.

For each stop we compute a 18-dimensional feature vector that counts various places of interests within a 200-meter radius of the stop; further tuning of the radius might improve the overall accuracy, but would be a side-issue for this paper. Two of the features are based on the bus data itself, indicating the number of other bus stops and tram stops within the area

Table 1: Average counts of the Foursquare venue categories within 200 meters of each stop.

Foursquare category	**Average**
Food	1.90
Home (private)	1.89
Shop & Service	1.77
Professional & Other Places	1.59
College & University	0.73
Nightlife Spot	0.61
Arts & Entertainment	0.48
Outdoors & Recreation	0.44
Residential Building	0.37
Bus Station	0.23
Hotel	0.19
Road	0.14
General Travel	0.02
Moving Target (food trucks etc.)	0.01
Train Station	0.01
Rental Car Location	0.01

(the tram routes are otherwise left out of the study), while the remaining 16 features are counts of different venue types crawled with the public Foursquare API, listed in Table 1. We do not use any other information about the venues besides the type and coordinates. The raw counts were preprocessed into three levels ("zero", "few" and "many"), by mean thresholding of the non-negative counts; we also tried simple logarithmic transformation that provided equivalent accuracy.

We want to model passenger volume for individual bus lines at different stops, and hence we treat all unique combinations of stops and routes (770 of them) as individual samples. The full covariate representation for each of these is 74-dimensional and consists of four parts: 18 dimensions to the feature vector of the stop, 18 dimensions corresponding to the weighted average of the feature vectors of other stops that can be reached by entering the bus, 37 dimensions corresponding to a binary encoding of the route identity, and finally a single dimension telling how far along the route the stop is (0=first stop, 1=the last stop).

The target features are estimated with a simple procedure that assumes people are less likely to do very short trips. We assume that the probability of getting out on each consecutive stop increases linearly until saturating after 5 stops. This part could be improved by learning the behavior from observed trips.

5.2 Experimental setup

The goal of our model is to predict transportation volume for route and stop combinations for which no training data is available, and hence the prediction power must come from the covariates. For this purpose we randomly split the data into training and test data in a stratified fashion that leaves roughly the same ra-

tio of stops along each route unobserved. Furthermore, we train the model using data of different days than what is being used for testing, to avoid using any part of the test data for training.

We train the model using data of only one day and test on the data of the same weekday during other weeks, using 385 of the 770 unique stop and route combinations for training and the rest for testing. This is effectively the hardest possible prediction setup for this application; we only observe the number of passengers during a single day on half of the stop-route -combinations and need to predict the passenger counts for stops for which we have no training data, purely based on the surroundings. We average results over 25 independent runs (5 random splits for each weekday) to make sure the findings are consistent.

We compare the two alternative variants (with and without the noise term ϵ_{nk}) of the proposed model (called LFR for *latent feature regression*) against five alternative strategies. The first two are univariate strategies applied independently for each output dimension, whereas the remaining three are multivariate solutions based on reduced-rank representation. For the samplers we discard the first 1000 samples as burn-in, and then sample for 1000 iterations keeping every 10th sample. The optimization-based models are learned until convergence.

1. LGNB regression model by Zhou et al. [2012], ran independently for each of the L dimensions.

2. Generalized linear model with Poisson likelihood with elastic net regularization, using the `glmnet` R package. This represents a standard univariate regression model that does not allow overdispersion but is regularized to prevent overfitting.

3. Reduced-rank vector generalized linear model (RR-VGLM) with Poisson likelihood, using the `VGAM` R package [Yee and Hastie, 2003]. We report the results for the best rank, which was 2; with 3 the method already severely overfits.

4. Gaussian latent feature regression as implemented in the `CCAGFA` R package using Bayesian formulation for canonical correlation analysis [Klami et al., 2013]. As pointed out by Breiman and Friedman [1997], CCA can be interpreted as deduced-rank regression, and its Bayesian variant is hence a fair comparison. We apply the model with and without log-transforming the data, rounding the predictions to get counts.

5. The DMR topic model [Mimno and McCallum, 2008] combined with LGNB regression [Zhou et al., 2012] for the total count of passengers for each sample. This represents one way of solving the problem with a more conventional topic modeling approach. It is worthwhile to note that the full model combining the topic model with the LGNB regression has not been presented before.

5.3 Results

We measure the error using two criteria. The first criterion is normalized mean square error (nMSE; normalized so that 1 corresponds to the variance of the data). It measures the overall accuracy, but ignores the uncertainty. The error is computed for the robust mean estimate (discarding the bottom and top 10% quantiles) of the predictive distributions.

The second criterion is predictive likelihood. To cope with potentially large counts (for which the probability is small for all methods), we measure the likelihoods for re-discretized counts with bins 0, 1, $2-3$, $4-7$, $8-15$, and so on. Furthermore, we report the accuracy for only the non-zero counts in the observed data, to avoid having the zeroes dominate the error measure (73% of the true counts are zero). For a good score in this measure, the method needs to assign high posterior probability for the right values and hence needs to also account for the variance of the predictions.

Figure 4 shows that in terms of mean prediction, RR-VGLM is the most accurate method followed by the two LFR variants. In terms of predictive likelihood the LRF models are by far the best, followed by CCA and LGNB+DMR, and hence best capture the whole predictive distribution. One notable observation is that both LFR variant are fragile for large K; even though the best likelihood is obtained with $K \geq 128$, the mean predictions are off for many folds due to rare cases with extremely large predictions caused by $p_{nk} \approx 1$. In practical use this could be avoided by truncation.

The combination of LGNB and DMR is conceptually close to LFR, but not as accurate. This is in part because learning the total count is hard; LGNB predicting directly the daily count has nMSE of 0.45 compared to 0.38 obtained by summing over the 20-minute predictions of our LFR model (with $K = 32$). LFR solves even this auxiliary problem of predicting the total daily count better than direct regression, by learning the sub-processes that explain the demand.

6 DISCUSSION

Even though count data is frequent in numerous applications and generalized linear models provide easy-to-use tools for modeling them, there has been surprisingly little work on multivariate regression on count data; we are only aware of vector generalized linear

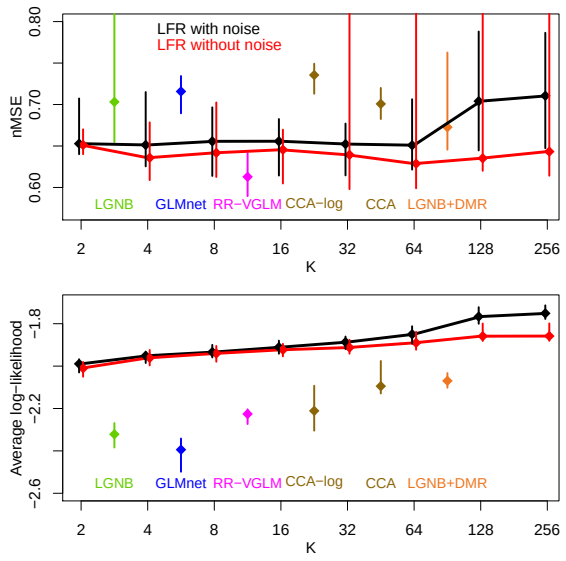

Figure 4: Performance on the public transportation data, shown as 10%, 50% and 90% quantiles over the 25 independent runs. The proposed method outperforms the comparisons clearly in log-likelihood (bottom plot), but RR-VGLM is more accurate in terms of the mean-square error (top plot). The lognormal noise variant (black line) has slightly better likelihood, but somewhat worse mean prediction compared to the other variant (red line; $\tau = \infty$). For LGNB+DMR we report the results for the best K; it is almost insensitive to this choice due to the LGNB part dominating the accuracy.

models by Yee and Wild [1996], their reduced-rank extensions [Yee and Hastie, 2003], and Bayesian solutions for low dimensionality [Ma et al., 2008].

To create a practical multivariate count regression method, we proposed a latent feature model that builds on recent Bayesian inference tools for negative binomial models [Zhou et al., 2012, Polson et al., 2013], and showed that it outperforms alternative solutions in modeling public transportation volume. The model could be further improved by using zero-inflation for the regression layer to improve sparsity and interpretability of the latent features. Reliable automatic method for selecting K would also be desirable; the priors on r and $\boldsymbol{\beta}$ prune out excess components (for $K = 256$ roughly half of the latent features become identically zero and the result is comparable to $K = 128$), but the large-K solutions are less robust than those with smaller K. For some folds they made large mistakes for some individual samples.

From another perspective, the model can be interpreted as an extension of the topic model conditioned on auxiliary features by Mimno and McCallum [2008]. Of particular interest is the idea of relaxing independence between document length and topic weights, which could be extended from the covariate-based model presented here to topic models in more general. Zhou and Carin [2012, 2015] touched the same issue by presenting topic models that assume negative binomial counts, but their inference strategy re-introduces the independence by treating the NB distribution as a mixture of Poisson distributions where the Poisson rates are explicitly instantiated.

We managed to capture 35% of the variance in passenger counts by regressing on a fairly simple static characterization, despite several simplifying assumptions like ignoring the schedules of the routes. For proper modeling of public transportation demand the predictive elements presented here need to be coupled with other kinds of models; the state-of-the-art in demand modeling today ignores the static surroundings.

Acknowledgements

The work was funded by the Academy of Finland (266969, 251170), TEKES (DIGILE D2I Programme), and the University Affairs Committee of the Xerox Foundation.

Appendix: Gibbs updates

The conditional distributions required for updating the regression part of (2) are provided by

$$
\begin{aligned}
(r_k|-) &\sim \mathrm{G}\Big(a_0 + \sum_{n=1}^{N} l_{nk}, \frac{1}{h_k + \sum_{n=1}^{N} \ln(1+e^{\psi_{nk}})}\Big),\\
(l_{nk}|-) &\sim \mathrm{CRT}(z_{nk}, r_k),\\
(h_k|-) &\sim \mathrm{G}(a_0 + b_0, 1/(e_0 + r_k)),\\
(\alpha_{dk}|-) &\sim \mathrm{G}(c_0 + 1/2, 1/(d_0 + \beta_{dk}^2/2)),\\
(\boldsymbol{\beta}_k|-) &\sim \mathrm{N}(\mathbf{m}_k, \mathbf{V}_k),\\
(\boldsymbol{\psi}_k|-) &\sim \mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k),\\
(\tau_k|-) &\sim \mathrm{G}\Big(f_0 + \frac{N}{2}, \frac{1}{g_0 + \|\boldsymbol{\psi}_k - \mathbf{X}\boldsymbol{\beta}_k\|_2^2/2}\Big),\\
(\omega_{nk}|-) &\sim \mathrm{PG}(z_{nk} + r_k, \psi_{nk}),
\end{aligned}
$$

where $\boldsymbol{\psi}_k = [\psi_{1k}, ..., \psi_{Nk}]^T$, $\boldsymbol{\Omega}_k = \mathrm{diag}(\omega_{1k}, ..., \omega_{Nk})$, $\mathbf{A}_k = \mathrm{diag}(\alpha_{1k}, ..., \alpha_{Dk})$, $\boldsymbol{\xi}_k = [(z_{1k} - r_k)/2, ..., (z_{Nk} - r_k)/2]^T$, $\mathbf{V}_k = (\tau_k \mathbf{X}^T\mathbf{X} + \mathbf{A}_k)^{-1}$, $\mathbf{m}_k = \tau_k \mathbf{V}_k \mathbf{X}^T \boldsymbol{\psi}_k$, $\boldsymbol{\Sigma}_k = (\tau_k \mathbf{I} + \boldsymbol{\Omega}_k)^{-1}$ and $\boldsymbol{\mu}_k = \boldsymbol{\Sigma}_k(\boldsymbol{\xi}_k + \tau_k \mathbf{X}\boldsymbol{\beta}_k)$. Finally, $\mathrm{CRT}(z_{nk}, r_k)$ denotes the distribution of the number of tables z_{nk} customers use in a Chinese restaurant process with concentration parameter r_k; see Zhou and Carin [2012] for details. In the alternative model where the lognormal noise element is omitted the parameters of the conditional of $\boldsymbol{\beta}_k$ are instead given by $\mathbf{V}_k = (\mathbf{X}^T\boldsymbol{\Omega}_k\mathbf{X} + \mathbf{A}_k)^{-1}$ and $\mathbf{m}_k = \mathbf{V}_k\mathbf{X}^T\boldsymbol{\xi}_k$.

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

Leo Breiman and Jerome H. Friedman. Predicting multivariate responses in multiple linear regression. *Journal of Royal Statistical Society B*, 59(3), 1997.

Gabriel Doyle and Charles Elkan. Accounting for burstiness in topic models. In *Proceedings of the 26th International Conference on Machine Learning*, 2009.

M.E. Ghitany, D. Karlis, D.K. Al-Mutairi, and F.A. Al-Awadhi. An EM algorithm for multivariate mixed Poisson regression models and its applications. *Applied Mathematical Sciences*, 6(137):6843–6856, 2012.

Javier Gutierrez, Osvaldo D. Cardozo, and Juan C. Garcia-Palomares. Transit ridership forecasting at station level: an approach based on distance-decay weighted regression. *Journal of Transport Geography*, 19(6):1081–1092, 2011.

Alan J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5:248–264, 1975.

Arto Klami. Pólya-gamma augmentations for factor models. In *Proceedings of the 6th Asian Conference on Machine Learning*, 2014.

Arto Klami, Seppo Virtanen, and Samuel Kaski. Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14:965–1003, 2013.

Jianming Ma, Kara M. Kockelman, and P. Damien. A multivariate poisson-lognormal regression model for prediction of crash counts by severity, using bayesian methods. *Accident Analysis and Prevention*, 40(3): 964–975, 2008.

David Mimno and Andrew McCallum. Topic models conditioned on arbitrary features with dirichlet-multinomial regression. In *Proceedings of Uncertainty in Artificial Intelligence*, 2008.

Nicholas G. Polson, James G. Scott, and Jesse Windle. Bayesian inference for logistis models using Polya-Gamma latent variables. *Journal of the American Statistical Association*, 108(504):1339–1349, 2013.

Thomas W. Yee and Trevor J Hastie. Reduced-rank vector generalized linear models. *Statistical Modelling*, 3:15–41, 2003.

Thomas W. Yee and Chris J. Wild. Vector generalized additive models. *Journal of the Royal Statistical Society B*, 58:481–493, 1996.

Ying Yuan, Yu Zheng, and Xing Xie. Discovering regions of different functions in a city using human mobility and POIs. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 186–194, 2012.

Mingyuan Zhou and Lawrence Carin. Augment-and-conquer negative binomial processes. In *Advances in Neural Information Processing Systems 25*, 2012.

Mingyuan Zhou and Lawrence Carin. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(2), 2015.

Mingyuan Zhou, Lingbo Li, David Dunson, and Lawrence Carin. Lognormal and gamma mixed negative binomial regression. In *Proceedings of the 29th International Conference on Machine Learning*, 2012.