Journal of Machine Learning Research ? (201?) ?-?? Submitted ??/??; Published ??/??

Generalized Nonparametric Deconvolution Models

Allison J.B. Chaney ACHANEY@CS.PRINCETON.EDU
Department of Computer Science
Princeton University
Princeton, NJ 08544, USA

Young-suk Lee YOUNGL@CS.PRINCETON.EDU
Department of Computer Science
Princeton University
Princeton, NJ 08544, USA

Barbara E. Engelhardt BEE@CS.PRINCETON.EDU
Department of Computer Science
Princeton University
Princeton, NJ 08544, USA

David M. Blei DAVID.BLEI@COLUMBIA.EDU
Department of Computer Science and Department of Statistics
Columbia University
New York, NY 10027, USA

Editor: TBD

Abstract

We describe *generalized nonparametric deconvolution models* (NDMs), a family of Bayesian nonparametric models for collections of data in which each observational unit is comprised of heterogeneous particles, such as in RNA sequencing. NDMs use the hierarchical Dirichlet process to discover the unknown number of latent factors that describe a dataset, and model each observation as a weighted average of these latent factors. Unlike existing models, however, NDMs are able to recover the factor-specific fluctuations for each observation. This allows us to deconvolve each observation into its constituent factors and describe how these factors deviate from their corresponding global factors. We present scalable variational inference techniques for this family of models and study its performance on RNA-seq, fMRI, image, financial, and demographic data. We show that including varied local factors improves prediction of missing data and yields interesting exploratory results.

Keywords: TBD

1. Introduction

We consider the problem of modeling collections of convolved data points. Specifically, each observation is composed of particles that originate from diverse factors. The objective of this work is to create a general family of models to learn 1) the features of global factors shared among all observations as well as the number and global proportions of these factors; 2) for

each observation, the proportion (or membership) of particles that belong to each factor; and 3) the features of observation-specific (or local) factors for each observation. While the first two objectives are fulfilled by existing models, the final objective is unique to our model.

Consider RNA sequencing data. Each observation is a cluster of cells that have been sequenced together—each cell has its own RNA, and their expressions are convolved together into a single observation. For a given cell sample, scientists would like to identify the proportions of different types of cells. For instance, a sample of blood cells will contain white blood cells, red blood cells, and platelets. In addition to learning the mixture proportions of a sample, scientists would also like to identify the RNA expression for each type of cell. (Some cells can be separated by type, but the process is expensive.) Finally, knowing the RNA expression *by type* and *for a specific sample* would be useful in determining which cell types are healthy or diseased for that sample.

This same structure exists in data from many disciplines, as shown in Table 1. For each of these domains, modeling local structure has distinct advantages, both in terms of predictive performance and interpretability. Because each discipline has its own nomenclature, we will use the general terminology of observations, which live in some high dimensional feature-space and are generated from some convolution of their constituent unobserved particles. It is our task to learn global and local factors to characterize the observations.

General	RNA-seq	fMRI	Photography	Financial	Demography
observation	(cell) sample	image	image	investment fund	regional population
feature	gene expression	voxel	pixel	stock	annual rate (e.g., birth)
particle	cell	neuron	light particle	investor	person
factor	cell type	response pattern	image feature	investment strategy	demographic pattern

Table 1: Structure of convolved observations in multiple domains

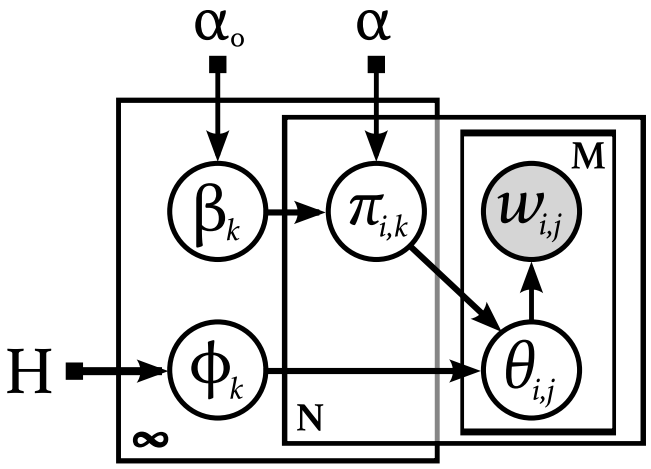

Figure 1: ■ HDP

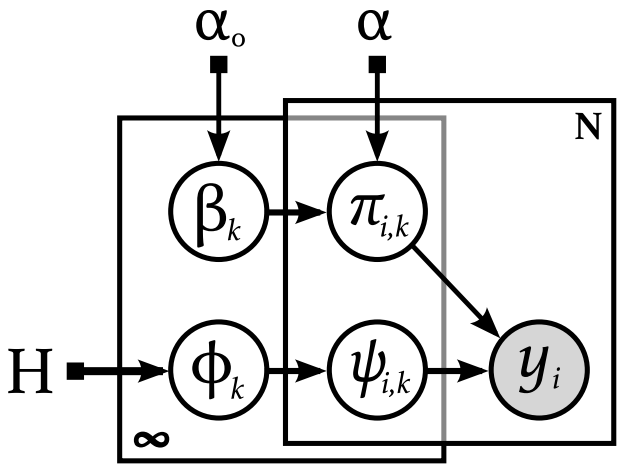

Figure 2: ■ ours

Related Work ¶ mixture models. What are they? (high level)

¶ admixture models.

¶ NMF

¶ other deconvolution models (mentioned below)

¶ density deconvolution

¶ HDP, in word (no math)

2. Nonparametric Deconvolution Models

In many disciplines, each observational unit is comprised of unobserved heterogeneous particles. In demography, for instance, countries are comprised of individual citizens, but often only country-level rates (e.g. birth rate) are reported. Each particle (e.g., citizen) can be represented in some high-dimensional feature space (e.g., income and education level), but can be summarized by a lower dimension of factors (e.g., socioeconomic status).

Just as individuals can be summarized by representative factors, the aggregate observations can also be summarized by these same factors—we can describe an observation in terms of the proportion of particles that belong to each factor. The United States, for example, could be described as having a population in which 50% of the citizens belong to the "middle class." The observations are then convolutions, or weighted averages, of the features of their constituent factors—observed average income and education levels in the United States would be weighted averages of socioeconomic class features.

Remember that investigators do not get to observe individual particles, but that the observation is an aggregation of its unobserved constituents. In order to learn the hidden factors that represent both observations and particles, we must then consider multiple convolved observations. This type of factorization or admixture analysis is common, and results in representing each observation in terms of global factors. Investigators, however, are interested not only in describing their data in terms of global patterns, but also in characterizing how observations deviate from these global patterns. In this section, we introduce a family of Bayesian nonparametric models that captures these local fluctuations; we term this family *generalized nonparametric deconvolution models* (NDMs).

■ move some of the above into the intro?

The NDM family is based on the hierarchical Dirichlet process (HDP) Teh et al. (2006), a Bayesian nonparametric model that describes the features of each sample as being generated from a mixture model; the mixture components are shared globally across all samples. We use the normalized gamma process construction of the HDP to formally specify the NDM family; we first review this construction then introduce NDMs.

Background: construction of the HDP The HDP is constructed using two layers of Dirichlet Processes (DPs). Following Paisley et al. (2012), we use two different representations

of the DP—one for each layer. The top-level DP uses the standard Sethuraman (1994) stick-breaking representation of the DP. As its name suggests, we imagine that some population can be broken down into its component parts much the way one would break a stick into pieces. Formally, we represent global proportions as β_k; this could describe, for example, how many middle class people there are as a percentage of the entire population. The Sethuraman generative process draws the unnormalized variant of these proportions β'_k from a beta distribution:

$$\beta'_k \,|\, \alpha_0 \sim \text{Beta}(1, \alpha_0). \tag{1}$$

The hyperparameter α_0 is called the concentration parameter and controls the distribution over the proportions. These proportions are normalized relative to all previous proportions,

$$\beta_k = \beta'_k \prod_{\ell=1}^{k-1} (1 - \beta'_\ell), \tag{2}$$

which gives us the stick breaking analogy: for the remainder of the population (or stick), how much should I assign to factor k (or how much of the stick should I break off)?

With β_k as our global proportions for factor k (how many middle class people are there?), we represent the features of factor k as ϕ_k (what features do middle class people usually have?). We generate these features from some base distribution H:

$$\phi_k \,|\, H \sim H. \tag{3}$$

When the HDP is used for topic modeling, the base distribution is usually a symmetric Dirichlet distribution over the feature simplex. For NDMs, H will take alternative forms.

The second layer of the HDP captures the same intuition as the top layer, but represents the local level instead of the global. The local proportions $\pi_{i,k}$ are the analog of the the global proportions β_k, with one set of proportions for each observation i. If the data is grouped by countries, then $\pi_{i,k}$ tells us what proportion of country i is made up of group k (e.g., the middle class). To generate these local proportions, we use a normalized gamma process Ferguson (1973), which begins by generating unnormalized proportions from a gamma distribution,

$$\pi'_{i,k} \,|\, \alpha, \beta_k \sim \text{Gamma}(\alpha\beta_k, 1), \tag{4}$$

and then normalizes them:

$$\pi_{i,k} = \frac{\pi'_{i,k}}{\sum_{\ell=1}^{\infty} \pi'_{i,\ell}}. \tag{5}$$

Like α_0, hyperparameter α is also a concentration parameter.

The NDM family uses this construction up until this point, when the models diverge. HDP models continue by using the local factor proportions π and the global factor features ϕ to construct discrete probability distributions G_i, one for each group of observations i:

$$G_i = \sum_{k=1}^{\infty} \pi_{i,k} \delta_{\phi_k}. \tag{6}$$

Each G_i is a distribution over the global factors ϕ_k—a draw from G_i produces ϕ_k with probability $\pi_{i,k}$.[1] Now we are able to draw local factor assignments θ,

$$\theta_{i,j} \,|\, G_i \sim G_i. \tag{7}$$

The observed data $w_{i,j}$ is then drawn from a distribution F parameterized by $\theta_{i,j}$:

$$w_{i,j} \,|\, \theta_{i,j} \sim F(\theta_{i,j}); \tag{8}$$

as the prior distribution H must be conjugate to F, F is usually multinomial.

As $\theta_{i,k}$ and $w_{i,k}$ do not have corresponding parameters in the NDM family, we have omitted demography analogies. Figure 1 shows a graphical model for the HDP.

NDM Generative Process As with the HDP construction, the NDM family draws global factor proportions β'_k (Eq. 1) and normalizes them to β_k (Eq. 2, "How much of the world is in the middle class?"). We also represent global factor features—"What attributes do middle class people usually have?"—with ϕ_k (Eq. 3).

At the local level, we also have factor proportions $\pi'_{i,k}$ (Eq. 4) which are similarly normalized to $\pi_{i,k}$ (Eq. 5, "How much of this country is in the middle class?"). Now deviating from the HDP construction, we draw *local factor features* $\psi_{i,k}$, which enable us to identify how local behavior or patterns deviate from global ones. With the demography example, instead of assuming that the middle class looks the same in every country, we can characterize how the middle class looks in each country individually—one country may have a more educated but lower household income relative to global patterns. Formally, we generate these local features

$$\psi_{i,k} \,|\, \phi_{i,k} \sim f(\phi_{i,k}), \tag{9}$$

where f is an arbitrary exponential family distribution parameterized by global factor features $\phi_{i,k}$. The choice of f encourages (but does not require) the selection of the base distribution H to be the corresponding conjugate prior; in section 3, we derive a generic inference algorithm that allows for non-conjugate relationships between f and H.

Given the local parameters, we then generate our observations y_i:

$$y_i \,|\, \pi_i, \psi_i \sim g\left(\sum_{k=1}^{\infty} \pi_{i,k}\psi_{i,k}\right), \tag{10}$$

where g is a generalized linear model. The entire generative process is captured by the graphical model in Figure 2.

NDM instances ¶ demography: f is poisson and g is ??; H is gamma?

¶ RNA: f is gaussian, g is linear regression (see NIPS); H is gaussian

¶ voxels (fMRI); f is gamma, logit function for g (logisitic regression??)?;

1. Other constructions simply draw an index to factor k with probability $\pi_{i,k}$. Either way, $\theta_{i,k}$ depends on both $\pi_{i,k}$ and ϕ_k—this construction just requires less bookeeping.

¶ images: f is gamma, g is linear??

¶ finance:

¶ explain specifics of related work after this point? (if we need to point out any mathematical details)

3. Inference

Our central computational problem is inference: given the observed data, how do we determine the best values for the latent parameters in our model? As the true posterior for our model is intractable to compute, we approach this problem with variational inference (Wainwright and Jordan, 2008).

Variational inference minimizes the KL divergence from an approximating distribution q to the true posterior p. This is equivalent to maximizing the ELBO:

$$\mathcal{L}(q) = \mathbb{E}_{q(\beta,\pi,\phi,\psi)}[\log p(y,\beta,\pi,\phi,\psi) - \log q(\beta,\pi,\phi,\psi)].$$

We define the approximating distribution q using the mean field assumption:

$$q(\beta,\pi,\phi,\psi) = \prod_{k=1}^{\infty} q(\beta'_k)q(\phi_k) \prod_{i=1}^{N} q(\pi'_{i,k})q(\psi_{i,k}),$$

where each latent variable-specific approximation q has the same form as its corresponding model distribution, but is parameterized by free variational parameters λ. For the distributions of factors, the global $q(\beta')$ is beta-distributed with variational parameter $\lambda^{\beta'}$, and the local $q(\pi')$ is gamma-distributed with variational parameter $\lambda^{\pi'}$. For the factor descriptions, global $q(\phi)$ has the same exponential family form as the base distribution H, and is parameterized by the variational parameter(s) λ^{ϕ}; local $q(\psi)$ has some exponential family form f, and is parameterized by λ^{ψ}.

To maximize the ELBO, we need to be able to compute the expectations of the hidden parameters under q. These expectations do not have a simple analytic forms, especially if we want to accommodate any settings of base distribution H from which we draw global factor features and the exponential family f from which we draw local factor features. To address the non-conjugate variables and accommodate a wide rage of settings for NDMs, we use "black box" variational inference techniques (Ranganath et al., 2015). Black box techniques sample from the variational distribution (e.g., $z_s \sim q(z \mid \lambda)$) to compute noisy unbiased gradients of the ELBO, then uses stochastic optimization to optimize the ELBO.

When we are updating an estimate of a latent variable, we only want to consider the parts of the posterior relevant to that parameter; this simplifies computation. We write the log probability of all terms containing a variable (its Markov blanket) for each of our latent variables:

$$\log p^{\beta'}(y,\beta,\pi,\phi,\psi) \triangleq \sum_{k=1}^{\infty} \left(\log p(\beta'_k \mid \alpha_0) + \sum_{i=1}^{N} \log p(\pi'_{i,k} \mid \alpha,\beta) \right), \tag{11}$$

$$\log p^{\pi'}(y, \beta, \pi, \phi, \psi) \triangleq \sum_{i=1}^{N} \left(\log p(y_i \mid \psi_i, \pi_i) + \sum_{k=1}^{\infty} \log p(\pi'_{i,k} \mid \alpha, \beta) \right), \tag{12}$$

$$\log p^{\phi}(y, \beta, \pi, \phi, \psi) \triangleq \sum_{k=1}^{\infty} \left(\log p(\phi_k \mid H) + \sum_{i=1}^{N} \log p(\psi_{i,k} \mid \phi_k) \right), \tag{13}$$

and

$$\log p^{\psi}(y, \beta, \pi, \phi, \psi) \triangleq \sum_{i=1}^{N} \left(\log p(y_i \mid \psi_i, \pi_i) + \sum_{k=1}^{\infty} \log p(\psi_{i,k} \mid \phi_k) \right). \tag{14}$$

We now write the gradients of each latent variable with respect to their corresponding variational parameters:

$$\nabla_{\lambda_k^{\beta'}} \mathcal{L} = \mathbb{E}_q \left[\nabla_{\lambda_k^{\beta'}} \log q(\beta'_k \mid \lambda_k^{\beta'}) \left(\log p^{\beta'}(y, \beta, \pi, \phi, \psi) - \log q(\beta'_k \mid \lambda_k^{\beta'}) \right) \right], \tag{15}$$

$$\nabla_{\lambda_{i,k}^{\pi'}} \mathcal{L} = \mathbb{E}_q \left[\nabla_{\lambda_k^{\pi'}} \log q(\pi'_k \mid \lambda_k^{\pi'}) \left(\log p^{\pi'}(y, \beta, \pi, \phi, \psi) - \log q(\pi'_k \mid \lambda_k^{\pi'}) \right) \right], \tag{16}$$

$$\nabla_{\lambda_k^{\phi}} \mathcal{L} = \mathbb{E}_q \left[\nabla_{\lambda_{i,k}^{\phi}} \log q(\phi_k \mid \lambda_k^{\phi}) \left(\log p^{\phi}(y, \beta, \pi, \phi, \psi) - \log q(\phi_k \mid \lambda_k^{\phi}) \right) \right], \tag{17}$$

and

$$\nabla_{\lambda_{i,k}^{\psi}} \mathcal{L} = \mathbb{E}_q \left[\nabla_{\lambda_{i,k}^{\psi}} \log q(\psi_{i,k} \mid \lambda_{i,k}^{\psi}) \left(\log p^{\psi}(y, \beta, \pi, \phi, \psi) - \log q(\psi_{i,k} \mid \lambda_{i,k}^{\psi}) \right) \right]. \tag{18}$$

We then use these gradients to construct our black box algorithm in Algorithm **??**.

■ note things like RMSprop and control variates...

■ explain sigmoid function $\mathcal{S}$ and P, h is family tht base distribtion H comes from... truncation?

■ walk through inference for several different settings of exponential family, and show table (like DEFs paper)

4. Empirical Results

¶ RNA, images, fMRI, demographic

¶ for each: introduce data, describe model parameterization, show exploratory results

¶ for simulated data: show that we recover ground truth?

Algorithm 1: Black Box Inference for Nonparametric Deconvolution Models

Input: observations y
Output: estimates of latent parameters β, π, ϕ, and ψ
Initialize variational parameters λ **Initialize** iteration count $t = 0$
while *change in validation likelihood* $< \delta$ **do**
  for *each sample* $s = 1, \ldots, S$ **do**
    for *each component* $k = 1, \ldots, K$ **do**
      draw sample global factor strength $\beta'_k[s] \sim \text{Beta}(\mathcal{S}(\lambda_k^{\beta'}))$
      draw sample global factor features $\phi_k[s] \sim h(\lambda_k^{\phi})$
      set $p_k^{\beta'}[s] = \log p(\beta'_k[s] \,|\, \alpha_0)$ // see Eqn. 1
      set $q_k^{\beta'}[s] = \log q(\beta'_k[s] \,|\, \lambda_k^{\beta'})$ // see Eqn. ??
      set $g_k^{\beta'}[s] = \nabla_{\lambda_k^{\beta'}} \log q(\beta'_k[s] \,|\, \lambda_k^{\beta'})$ // see Eqn. ??
      set $p_k^{\phi}[s] = \log p(\phi_k[s] \,|\, \alpha_0)$ // see Eqn. 3
      set $q_k^{\phi}[s] = \log q(\phi_k[s] \,|\, \lambda_k^{\phi})$ // see Eqn. ??
      set $g_k^{\phi}[s] = \nabla_{\lambda_k^{\phi}} \log q(\phi_k[s] \,|\, \lambda_k^{\phi})$ // see Eqn. ??
    end
    for *each observation* $i = 1, \ldots, N$ **do**
      for *each component* $k = 1, \ldots, K$ **do**
        draw sample local factor strength $\pi'_{i,k}[s] \sim \text{Gamma}(\mathcal{P}(\lambda_{i,k}^{\pi'}))$
        draw sample local factor features $\psi_{i,k}[s] \sim f(\lambda_{i,k}^{\psi})$
        set $p_{i,k}^{\pi'}[s] = \log p(\pi'_{i,k}[s] \,|\, \alpha, \beta)$ // see Eqn. ??
        set $q_{i,k}^{\pi'}[s] = \log q(\pi'_{i,k}[s] \,|\, \lambda_{i,k}^{\pi'})$ // see Eqn. ??
        set $g_{i,k}^{\pi'}[s] = \nabla_{\lambda_{i,k}^{\pi'}} \log q(\pi'_{i,k}[s] \,|\, \lambda_{i,k}^{\pi'})$ // see Eqn. ??
        set $p_{i,k}^{\psi}[s] = \log p(\psi_{i,k}[s] \,|\, \phi)$ // see Eqn. ??
        set $q_{i,k}^{\psi}[s] = \log q(\psi_{i,k}[s] \,|\, \lambda_{i,k}^{\psi})$ // see Eqn. ??
        set $g_{i,k}^{\psi}[s] = \nabla_{\lambda_{i,k}^{\psi}} \log q(\psi_{i,k}[s] \,|\, \lambda_{i,k}^{\psi})$ // see Eqn. ??
      end
      set $p_i^y[s] = \log p(y_i \,|\, g\left(\sum_{k=1}^{i} nfty\pi_{i,k}\psi_{i,k}\right))$ // see Eqn. 1
    end
  end
  set $\rho = (t + \tau)^{\kappa}$
  for *each component* $k = 1, \ldots, K$ **do**
    set $\hat{\nabla}_{\lambda_k^{\alpha}} \mathcal{L} \triangleq \frac{1}{S} \sum_{s=1}^{S} g_k^{\alpha}[s](p_k[s] - q_k[s])$
    set $\hat{\nabla}_{\lambda_k^{\mu}} \mathcal{L} \triangleq \frac{1}{S} \sum_{s=1}^{S} g_k^{\mu}[s](p_k[s] - q_k[s])$
    set λ^{α} += $\rho \hat{\nabla}_{\lambda_k^{\alpha}} \mathcal{L}$
    set λ^{μ} += $\rho \hat{\nabla}_{\lambda_k^{\mu}} \mathcal{L}$
  end
end
for *each component* $k = 1, \ldots, K$ **do**
  set $\mathbb{E}[\mu_k] = \lambda_k^{\mu}$
end
return $\mathbb{E}[\mu]$

5. Discussion

¶ Not sure what will go here...summary of contributions and results?

Acknowledgments

We would like to acknowledge ...

References

Thomas S Ferguson. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pages 209–230, 1973.

John Paisley, Chong Wang, David M Blei, et al. The discrete infinite logistic normal distribution. *Bayesian Analysis*, 7(4):997–1034, 2012.

Rajesh Ranganath, Linpeng Tang, Laurent Charlin, and David M. Blei. Deep exponential families. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, AISTATS '15, pages 762–771, 2015.

Jayaram Sethuraman. A constructive definition of dirichlet priors. *Statistica Sinica*, 4: 639–650, 1994.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the American statistical association*, 101(476), 2006.

Martin J. Wainwright and Michael I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1-2):1–305, January 2008. ISSN 1935-8237. doi: 10.1561/2200000001. URL http://dx.doi.org/10.1561/2200000001.

Appendix A. Whatever it is called.

¶ most likely some inference details will go here