The third assumption is the *a priori* choice of the number of topics. The most direct nonparametric extension of LDA is the *hierarchical Dirichlet process* (HDP) [17]. The HDP allows an unbounded set of topics via a latent stochastic process, but nevertheless imposes a Dirichlet distribution on any finite subset of these topics. Alternatively, the *nonparametric Bayes pachinko allocation* [9] model captures correlations within an unbounded topic collection via an inferred, directed acyclic graph. More recently, the *discrete infinite logistic normal* [13] (DILN) model of topic correlations used an exponentiated Gaussian process (GP) to rescale the HDP. This construction is based on the gamma process representation of the DP [5]. While our goals are similar, we propose a rather different model based on the stick-breaking representation of the DP [16]. This choice leads to arguably simpler learning algorithms, and also facilitates our modeling of document metadata.

In this paper, we develop a *doubly correlated nonparametric topic* (DCNT) model which captures between-topic correlations, as well as between-document correlations induced by metadata, for an unbounded set of potential topics. As described in Sec. 2, the global soft-max transformation of the DMR and CTM is replaced by a stick-breaking transformation, with inputs determined via both metadata-dependent linear regressions and a square-root covariance representation. Together, these choices lead to a well-posed nonparametric model which allows tractable MCMC learning and inference (Sec. 3). In Sec. 4, we validate the model using a toy dataset, as well as a corpus of NIPS documents annotated by author and year of publication.

2 A Doubly Correlated Nonparametric Topic Model

The DCNT is a hierarchical, Bayesian nonparametric generalization of LDA. Here we give an overview of the model structure (see Fig. 1), focusing on our three key innovations.

2.1 Document Metadata

Consider a collection of D documents. Let $\phi_d \in \mathbb{R}^F$ denote a feature vector capturing the metadata associated with document d, and ϕ an $F \times D$ matrix of corpus metadata. When metadata is unavailable, we assume $\phi_d = 1$. For each of an unbounded sequence of topics k, let $\eta_{fk} \in \mathbb{R}$ denote an associated significance weight for feature f, and $\eta_{:k} \in \mathbb{R}^F$ a vector of these weights.[2]

We place a Gaussian prior $\eta_{:k} \sim N(\mu, \Lambda^{-1})$ on each topic's weights, where $\mu \in \mathbb{R}^F$ is a vector of mean feature responses, and Λ is an $F \times F$ diagonal precision matrix. In a hierarchical Bayesian fashion [6], these parameters have priors $\mu_f \sim N(0, \gamma_\mu)$, $\lambda_f \sim \mathrm{Gam}(a_f, b_f)$. Appropriate values for the hyperparameters γ_μ, a_f, and b_f are discussed later.

Given η and ϕ_d, the document-specific "score" for topic k is sampled as $u_{kd} \sim N(\eta_{:k}^T \phi_d, 1)$. These real-valued scores are mapped to document-specific topic frequencies π_{kd} in subsequent sections.

2.2 Topic Correlations

For topic k in the ordered sequence of topics, we define a sequence of k linear transformation weights $A_{k\ell}$, $\ell = 1, \ldots, k$. We then sample a variable v_{kd} as follows:

$$v_{kd} \sim N\left(\sum_{\ell=1}^{k} A_{k\ell} u_{\ell d}, \lambda_v^{-1}\right) \tag{1}$$

Let A denote a lower triangular matrix containing these values $A_{k\ell}$, padded by zeros. Slightly abusing notation, we can then compactly write this transformation as $v_{:d} \sim N(Au_{:d}, L^{-1})$, where $L = \lambda_v I$ is an infinite diagonal precision matrix. Critically, note that the distribution of v_{kd} depends only on the first k entries of $u_{:d}$, not the infinite tail of scores for subsequent topics.

Marginalizing $u_{:d}$, the covariance of $v_{:d}$ equals $\mathrm{Cov}[v_{:d}] = AA^T + L^{-1} \triangleq \Sigma$. As in the classical factor analysis model, A encodes a square-root representation of an output covariance matrix. Our integration of input metadata has close connections to the semiparametric latent factor model [18], but we replace their kernel-based GP covariance representation with a feature-based regression.

[2] For any matrix η, we let $\eta_{:k}$ denote a column vector indexed by k, and $\eta_{f:}$ a row vector indexed by f.