Variational Bayes for Merging Noisy Databases

Tamara Broderick
University of California, Berkeley
tab@stat.berkeley.edu

Rebecca C. Steorts
Carnegie Mellon University
beka@cmu.edu

October 20, 2014

Abstract

Bayesian entity resolution merges together multiple, noisy databases and returns the minimal collection of unique individuals represented, together with their true, latent record values. Bayesian methods allow flexible generative models that share power across databases as well as principled quantification of uncertainty for queries of the final, resolved database. However, existing Bayesian methods for entity resolution use Markov monte Carlo method (MCMC) approximations and are too slow to run on modern databases containing millions or billions of records. Instead, we propose applying variational approximations to allow scalable Bayesian inference in these models. We derive a coordinate-ascent approximation for mean-field variational Bayes, qualitatively compare our algorithm to existing methods, note unique challenges for inference that arise from the expected distribution of cluster sizes in entity resolution, and discuss directions for future work in this domain.

1 Introduction

Merging records from multiple databases is a problem that emerged from the genetics literature [13] and is a pressing issue in statistics and computer science in the modern day [5]. For instance, human rights organizations collect records of war crimes in the Middle East and Central America and want to estimate the total number of victims [11]. The United States Census Bureau wants to estimate minority representation and child poverty in different parts of the country [6, 1]. In each of these examples, individual records are collected in multiple databases. Due to the collection procedure, records are often duplicated within a single database and across databases. Crucially, due to various factors, some records in these databases are corrupted by noise. In any case, an important part of delivering an estimate for any quantity of interest from the merged databases is also returning some uncertainty for that estimate. [15, 16, 17, 14] have recently applied a Bayesian statistical paradigm to merging databases by modeling the noisy corruption as a random process; the authors have shown that their approach provides not only desirable uncertainty quantification for a variety of model queries but also flexible generative models to capture the many unique types of records and record relationships that may be present in these databases. However, the MCMC approximations used in these Bayesian analyses do not scale sufficiently to process the large number of records in many modern and complex databases. Thus, we propose a variational Bayes approximation to capture desired uncertainty in posterior Bayesian estimates while simultaneously allowing the processing of much larger and more realistic databases than is possible with these methods. Finally, we elaborate on how these database-merging models pose unique challenges for variational approximations.

2 Background

Entity resolution refers to the merging of multiple databases (often without shared unique identifiers) into a single database of unique entities [5]. Special cases of entity resolution include *record linkage*, which refers to the identification of records across different databases that represent the same entity, and *de-duplication*, which refers to the identification of records within the same database that represent the same entity. Traditional approaches for entity resolution that link records directly to other records become computationally infeasible as the number of records grows [5, 19]. Here, we instead take the approach of [17] and imagine each record as representing a latent individual. When we further suppose that some entries may suffer from noisy corruption, entity resolution can be viewed as a clustering problem. The observed data being clustered are the records in each database, and the latent cluster centers are the unobserved, latent individuals.

It is common for record fields to be discrete or categorical: e.g., county of residence, race, gender, etc. Like [16], we focus on categorical data in what follows.[1] [3, 2] previously demonstrated the scaling advantages of variational Bayesian approximations to posteriors for mixture and admixture modeling, where the observed data are categorically-valued (e.g. words in a vocabulary). There exist unique challenges in our data clustering problem. A particularly important one lies in an assumption inherent in many of the popular Bayesian models for clustering and admixture—such as mixture models, LDA [3], Dirichlet processes [12], hierarchical Dirichlet processes [18], and many more. These models all implicitly assume that any cluster makes up a non-zero proportion of the data that does not change as the data set size increases without bound [9, 4]. This assumption, by contrast, is very clearly inappropriate for entity resolution problems. In entity resolution, we expect each cluster to contain perhaps one and at most a handful of records. We expect the number of clusters to grow linearly with the number of records in a given database—though the number of clusters might be constant as more databases are added. We call the problem of modeling this type of clustering behavior, which differs from classical models and assumptions, the *small clustering problem*, and we discuss it in more detail in Section 4.

[1] The nature of data corruption in the database collection process can often lead to interesting and non-standard noise distributions for data that might, in other contexts, be treated as non-categorical or continuous (date of birth, age, etc.). Text fields—e.g., name of an individual—must also be treated with more care than assigning a single categorical distribution in this context. These considerations, though addressed elsewhere [14, 15], are outside the scope of this note.

3 A new generative model for entity resolution

Let D represent the number of databases and R_d represents the number of records in the dth database. All records contain the same F fields, which are categorical. Let field f have V_f possible values, or *field attributes*. Assume for now that every record is complete; that is, for every database and record, there is no missing field. Let x_{drf} be the observed data value in the fth field of the rth record in the dth database. We make the further simplifying assumption that there are K unique latent individuals. Ultimately, we desire a model where K is random, and we learn a posterior distribution over K given the full data $x = \{x_{drf}\}_{d,r,f}$ across *all* databases, records, and fields. But as a first step we assume K is fixed and known, as in LDA [3]. Let z_{dr} be the latent individual for whom (potentially noisy) data $x_{dr\cdot}$ is recorded in the rth record of the dth database.

In other words, we regard each record $x_{dr\cdot}$ as a possibly distorted copy of an ideal latent record for latent individual z_{dr}. To capture this idea, let $\beta_{kf\cdot}$ be a discrete noise distribution associated with the kth latent individual. That is, β_{kfv}, $v \in \{1, \dots, V_f\}$, are numbers between zero and one that sum to one across v. If there were no noise in the data entry procedure, the probabilities $\beta_{kf\cdot}$ would correspond to a trivial distribution with all of its mass at some true latent value v^*_{kf} of the fth field for the kth individual: $\beta_{kfv} = \mathbb{1}\{v = v^*_{kf}\}$.[2] In general, there is some noise in the records, and β_{kfv} corresponds to a non-trivial noise distribution; however we assume it has a plurality of its mass at the true value.

For our generative model, assume that the observed value x_{drf} of the fth field in record r in database d is drawn from the noise distribution associated with the latent individual z_{dr} for this record; that is,

$$x_{drf} | \beta_{\cdot f \cdot}, z_{dr} \sim \text{Categorical}_{V_f}(\beta_{z_{dr} f \cdot}),$$

where Categorical_{V_f} is the categorical distribution over $1, \dots, V_f$ with probabilities given by the distribution parameter. These draws are independent across records and fields, conditional on β and z.

Next, we form a hierarchical Bayesian model by putting priors on both z and β. For z, we assume that the latent individual for any record is drawn uniformly over all latent individuals and independently across records:

$$z_{dr} \sim \text{Categorical}_K(K^{-1}\, \mathbf{1}_K),$$

where $\mathbf{1}_K$ is the vector of all ones of length K. Since $\beta_{kf\cdot}$ is a vector of probabilities, a natural choice of prior for $\beta_{kf\cdot}$ is the Dirichlet distribution on a vector of size V_f, which we denote by Dirichlet_{V_f}. Thus, we assume that the $\beta_{kf\cdot}$ vectors are drawn independently according to

$$(\beta_{kfv})_{v=1}^{V_f} \sim \text{Dirichlet}_{V_f}(A_\cdot),$$

with hyperparameter vector $A_\cdot = (A_1, \dots, A_{V_f})$. Typically, we assume that the A_v are small (near zero) so that the Dirichlet parameter encourages $\beta_{kf\cdot}$ to be peaked around a single value. We typically choose $A_1 = \dots = A_{V_f}$.

[2]Here, $\mathbb{1}(E)$ is the indicator function for event E.

4 Comparison with previous work

We briefly review the model of [17], where the authors introduced the basic Bayesian clustering framework for entity resolution and their Split and MErge REcord linkage and De-duplication (SMERED) algorithm. [17] took a fully hierarchical-Bayesian approach, in the special case where all the record fields are categorical and independent. The authors derived an efficient hybrid (Metropolis-within-Gibbs) MCMC algorithm, SMERED. SMERED is able to update most of the latent variables and parameters using Gibbs sampling steps from conjugate conditional distributions. While SMERED updates the assignment of records to latent individuals using a split-merge step, following [8], and can run on a health care databases of 60,000 records in 3.5 hours, it does not scale to "large databases."[3] In terms of scalability, we wish to scale to millions or billions of records in one or multiple databases. For example, the U.S. Census contains approximately 300 million records, while many medical databases at large universities or in the entire country would contain millions or billions of records.

Furthermore, while the model of [17] was shown to work very well for entity resolution applications, it is not easily approximated with variational methods due to various deterministic dependencies in the generative model. We show the full model for SMERED in Appendix A, where we also provide a mapping between the SMERED model and our new generative model from Section 3. By contrast, we directly demonstrate in Section 5 how our new generative model, which is inspired by LDA, is readily amenable to variational approximation.

While our model has some similarities to LDA, there are also some differences—large and small. For one, the fields do not enjoy the symmetry of words in a bag-of-words model of a document; that is, the fields are ordered and cannot be interchanged. Second, as we do not expect the distribution of individuals to vary wildly by database, we keep the same uniform distribution over latent individuals in each database. By contrast, an important part of LDA is allowing the admixture proportions of topics to vary by document.

We now raise a key issue regarding the main distinction between many classical Bayesian models for mixtures and admixtures (such as mixture models, LDA, feature-allocation models [4] including the Indian buffet process [7], etc.) and entity resolution. The issue arises from framing entity resolution as a clustering problem. In clustering and other statistical models, it is common to assume that our data are *infinitely exchangeable*, meaning that for any data set size, we assume that the distribution of our data would not change if the data were observed in a different order. This simple assumption applied to clustering models implies, via the *Kingman paintbox* [9, 4], that every cluster forms some strictly positive proportion of the data, and this proportion does not change as the data grows. In mixture models, these are the mixing proportions; there may be finitely many in a finite mixture model or infinitely many in a Dirichlet process model. In any of these cases, there are two important consequences for our model. First, as the data set size grows, we always observe more data points in a

[3] This database is the National Long Term Care Study (NLTCS), a longitudinal study of the health status of elderly Americans `http://www.nltcs.aas.duke.edu/`. The authors ran the NLTCS on three databases of 20,000 records each, for 1 million iterations of their hybrid MCMC, which took 3.5 hours to run.

cluster. In fact, the number of observed data in a cluster grows without bound. Second, because the size of every cluster grows to infinity as the data set size grows to infinity, the usual asymptotic theory applies to inferring cluster properties or parameters. Uncertainty about, e.g., a cluster mean typically shrinks to zero in the limit.

When clusters are unique individuals in a population, however, it is not natural to assume that more data always eventually means more records of the same individual. Rather, every cluster should be observed a strictly finite number of times. This means that uncertainty about latent individuals cannot shrink to zero (in general). Since the assumptions of the traditional models (such as LDA) are violated in this case, they do not apply. And we must ask: what are natural regularity assumptions in this *small clustering* domain, what inferences can we draw about clusters in this domain, and what new families of distributions can we apply? A similar issue to what we have dubbed the small clustering problem has previously been identified for infinitely exchangeable graphs by [10]. This problem is also reminiscent of challenges in high-dimensional statistics, where the number of parameters may grow linearly (or much faster) than the data size.

5 Mean-field variational approximation

The generative model specified in Section 3 yields the following joint distribution for data x and parameters β, z:

$$p(\beta, z, x) = \left[\prod_{k=1}^{K} \prod_{f=1}^{F} \text{Dirichlet}_{V_f}(\beta_{kf\cdot} | A_\cdot) \right] \left[\prod_{d=1}^{D} \prod_{r=1}^{R_d} \prod_{f=1}^{F} \beta_{z_{dr} l x_{drf}} \right]. \tag{1}$$

Note that the posterior on the parameters, $p(\beta, z|x)$, is proportional to $p(\beta, z, x)$.

As this posterior cannot be solved for in closed form, we must approximate it. Here, we consider a variational approximation q of the following form:

$$q(\beta, z) = \left[\prod_{d=1}^{D} \prod_{r=1}^{R_d} q(z_{dr} | \phi_{dr\cdot}) \right] \left[\prod_{k=1}^{K} \prod_{f=1}^{F} q(\beta_{kf\cdot} | \lambda_{kf\cdot}) \right], \tag{2}$$

where we have introduced variational parameters ϕ and λ. We further assume

$$q(z_{dr} | \phi_{dr\cdot}) = \text{Categorical}_K(\phi_{dr\cdot}) \quad \text{and} \quad q(\beta_{kf\cdot} | \lambda_{kf\cdot}) = \text{Dirichlet}_{V_f}(\beta_{kf\cdot} | \lambda_{kf\cdot}). \tag{3}$$

The variational optimization problem is to minimize the Kullback-Leibler divergence from $q(\beta, z)$ to $p(\beta, z|x)$: $\min_{\lambda,\phi} \text{KL}\left(q_{\lambda,\phi}(\beta, z) \,\|\, p(\beta, z|x)\right)$. To clarify that the optimization is over different choices of the distribution q, which are indexed by parameters λ and ϕ, we write $q(\beta, z)$ as $q_{\lambda,\phi}(\beta, z)$ above. We derive the following coordinate-ascent steps in the variational parameters ϕ and λ for the variational optimization problem $\min_{\lambda,\phi} \text{KL}(q_{\lambda,\phi} \| p)$ in Appendix B:

$$\lambda_{kfv} \leftarrow A_v + \sum_{d=1}^{D} \sum_{r=1}^{R_d} \phi_{drk} \mathbb{1}\{x_{drf} = v\},$$

$$\phi_{drk} \propto_k \exp\left\{\sum_{f=1}^{F}\sum_{v=1}^{V_f}\mathbb{1}\{x_{drf}=v\}\left[\psi(\lambda_{kfv})-\psi\left(\sum_{u=1}^{V_f}\lambda_{kfv}\right)\right]\right\}.$$

6 Future directions

Our next step is to compare the algorithm resulting from our new generative model and variational approximation to the existing Bayesian model and resulting MCMC algorithm SMERED of [17]. We anticipate the variational approach will be much faster, however, there may be accuracy tradeoffs. Also, since we have made many simplifying assumptions, we propose incorporation of more realistic assumptions about record fields as in [14, 15]. Moreover, it remains to allow the number of latent individuals to grow with the size of the data, to construct a model that allows posterior inference of this number, and to address the small clustering problem we have posed. We wish to find a solution that addresses this problem not only for entity resolution but more broadly in other domains, where clusters may not be expected to grow without bound as a proportion of the total data.

7 Acknowledgements

TB was supported by the Berkeley Fellowship. RCS was supported by NSF grants SES1130706 and DMS1043903 and NIH grant #1 U24 GM110707-01.

References

[1] W.R. Bell, G.S. Datta, and M. Ghosh. Benchmarked small area estimators. *Biometrika*, 100(1):189–202, 2013.

[2] D. M. Blei and M. I. Jordan. Variational inference for Dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 2006.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] T. Broderick, J. Pitman, and M. I. Jordan. Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, 8(4):801–836, 2013.

[5] P. Christen. *Data Matching: Concepts and Techniques for Record Linkage, Entity Resolution, and Duplicate Detection*. Springer, 2012.

[6] G. S. Datta, M. Ghosh, R. Steorts, and J. Maples. Bayesian benchmarking with applications to small area estimation. *TEST*, 20(3):574–588, 2011.

[7] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the Indian buffet process. Technical report, TR2005-001, Gatsby Unit, 2005.

[8] S. Jain and R. Neal. A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model. *Journal of Computational and Graphical Statistics*, 13:158–182, 2004.

[9] J. F. C. Kingman. The representation of partition structures. *Journal of the London Mathematical Society*, 2(2):374–380, 1978.

[10] J. Lloyd, P. Orbanz, Z. Ghahramani, and D. Roy. Random function priors for exchangeable arrays with applications to graphs and relational data. In *NIPS*, 2012.

[11] Kristian Lum, Megan Emily Price, and David Banks. Applications of multiple systems estimation in human rights research. *The American Statistician*, 67(4):191–200, 2013.

[12] S. N. MacEachern and P. Müller. Estimating mixture of Dirichlet process models. *Journal of Computational and Graphical Statistics*, 7(2):223–238, 1998.

[13] H. B. Newcombe, J. M. Kennedy, S. J. Axford, and A. P. James. Automatic linkage of vital records computers can be used to extract" follow-up" statistics of families from files of routine records. *Science*, 130(3381):954–959, 1959.

[14] M. Sadinle. Detecting duplicates in a homicide registry using a Bayesian partitioning approach. *Annals of Applied Statistics*, *forthcoming*, 2014. http://arxiv.org/abs/1407.8219.

[15] R. C. Steorts. Empirical Bayesian graphical entity resolution. *Submitted*, 2014. http://arxiv.org/abs/1409.0643.

[16] R. C. Steorts, R. Hall, and S. E. Fienberg. A Bayesian approach to graphical record linkage and de-duplication. *In Revision*, 2014. http://arxiv.org/abs/1312.4645.

[17] R. C. Steorts, R. Hall, and S. E. Fienberg. SMERED: A Bayesian approach to graphical record linkage and de-duplication. *Journal of Machine Learning Research*, 33:922–930, 2014. http://arxiv.org/abs/1403.0211.

[18] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476), 2006.

[19] W. E. Winkler. Overview of record linkage and current research directions. In *Bureau of the Census*. Citeseer, 2006.

A Review of SMERED and notational map with new generative model

The independent fields model of [17] assumes the d databases are conditionally independent, given the latent individuals, and that fields are independent within individuals. We use the same notation as the generative model earlier, with D databases, R_d records within the dth database, and F fields within each record. Then $\boldsymbol{x}_{dr}$ is a categorical vector of length p. Let $\boldsymbol{y}_k$ be the latent vector of true field values for the kth record, where $k \in \{1, \ldots, K\}$ indexes the latent individuals. The *linkage structure* is defined as $\boldsymbol{\Lambda} = \{\lambda_{dr} \ ; \ d = 1, \ldots, D \ ; \ r = 1, \ldots, R_d\}$ where λ_{dr} is an integer from 1 to K indicating the latent individual to which the rth record in database d refers, i.e., $\boldsymbol{x}_{dr}$ is a possibly-distorted measurement of $\boldsymbol{y}_{\lambda_{dr}}$. Finally, $\tilde{z}_{drf}$ is 1 or 0 according to whether or not the particular field f is distorted in $\boldsymbol{x}_{dr}$.

The Bayesian parametric model is

$$\boldsymbol{x}_{drf} \mid \lambda_{dr}, \boldsymbol{y}_{\lambda_{dr}\ell}, z_{drf}, \boldsymbol{\theta}_f \overset{\text{ind}}{\sim} \begin{cases} \delta_{\boldsymbol{y}_{\lambda_{dr}f}} & \text{if } z_{drf} = 0 \\ \text{Categorical}(1, \boldsymbol{\theta}_f) & \text{if } z_{drf} = 1 \end{cases} \tag{4}$$

$$\tilde{z}_{drf} \overset{\text{ind}}{\sim} \text{Bern}(\tilde{\beta}_f)$$

$$\boldsymbol{y}_{kf} \mid \boldsymbol{\theta}_f \overset{\text{ind}}{\sim} \text{Categorical}(1, \boldsymbol{\theta}_f) \tag{5}$$

$$\boldsymbol{\theta}_f \overset{\text{ind}}{\sim} \text{Dirichlet}(\boldsymbol{\mu}_f) \tag{6}$$

$$\tilde{\beta}_f \overset{\text{ind}}{\sim} \text{Beta}(a_f, b_f) \tag{7}$$

$$\pi(\boldsymbol{\Lambda}) \propto 1. \tag{8}$$

To compare the SMERED generative model to our generative model in Section 3, first note that the observed rth record in database d is $x_{dr\cdot}$ in both models. The latent individual for the record at (d, r) in SMERED is λ_{dr} and in Section 3 it is z_{dr}. In SMERED, the noise distribution for a latent individual is separated into two steps: whether there is noise for a given field ($\tilde{z}_{drf}$) and the distribution of that noise ($\boldsymbol{\theta}_f$). By contrast, in Section 3, $\beta_{kf\cdot}$ captures the full distribution of field values for individual k. Also, while SMERED places a distribution $\boldsymbol{\theta}_f$ on the underlying distribution of field values, such a distribution is implicit in aggregating over $\beta_{kf\cdot}$ in Section 3. Likewise, the "true record values" of SMERED's $\boldsymbol{y}_{\lambda_{dr}\cdot}$ are implicit in the distribution $\beta_{z_{dr}f\cdot}$ of Section 3.

B Mean-field variational approximation derivation

B.1 Mean-field variational problem

We recall that minimizing the Kullback Leibler divergence (KL) divergence,

$$\min_{\lambda, \phi} \text{KL}(q_{\lambda,\phi}(\beta, z) || p(\beta, z | x)),$$

is equivalent to maximizing $\mathrm{ELBO}(\phi,\lambda)$, where

$$\begin{aligned}\mathrm{ELBO}(\phi,\lambda) &= -\mathrm{KL}(q_{\lambda,\phi}(\beta,z)||p(\beta,z|x)) + p(x)\\ &= \mathbb{E}_q[\log p(\beta,z,x)] - \mathbb{E}_q[\log q(\beta,z;\phi,\lambda)].\end{aligned}$$

Henceforth, we concentrate on maximizing $\mathrm{ELBO}(\phi,\lambda)$ with respect to mean-field approximation parameters ϕ,λ.

From the generative model in Section 3, we derived the joint distribution of parameters β,z and data x, $p(\beta,z,x)$, in Eq. (1). We also assume that the approximating distribution $q(\beta,z)$ for the posterior $p(\beta,z|x)$ takes the form specified in Eqs. (2) and (3). Using these equations, we find

$$\begin{aligned}\mathrm{ELBO}(\phi,\lambda) = &\sum_{k=1}^{K}\sum_{f=1}^{F}\mathbb{E}_q[\log \mathrm{Dirichlet}_{V_f}(\beta_{kf\cdot}|A)]\\ &+\sum_{d=1}^{D}\sum_{r=1}^{R_d}\sum_{f=1}^{F}\sum_{k=1}^{K}\sum_{v=1}^{V_f}\mathbb{E}_q[\mathbb{1}\{z_{dr}=k\}\mathbb{1}\{x_{drf}=v\}\log(\beta_{kfv})]\\ &-\sum_{d=1}^{D}\sum_{r=1}^{R_d}\sum_{k=1}^{K}\mathbb{E}_q[\mathbb{1}\{z_{dr}=k\}\log(\phi_{drk})]\\ &-\sum_{k=1}^{K}\sum_{f=1}^{F}\mathbb{E}_q[\log \mathrm{Dirichlet}_{V_f}(\beta_{kf\cdot}|\lambda_{kf\cdot})].\end{aligned}$$

To evaluate these expectations, we recall the definitions of the *digamma* ψ and *trigamma* functions ψ_1:

$$\begin{aligned}\psi(x) &= \frac{d}{dx}\log\Gamma(x)\\ \psi_1(x) &= \frac{d^2}{dx^2}\log\Gamma(x) = \frac{d}{dx}\psi(x).\end{aligned}$$

With these functions in hand, we can write

$$\begin{aligned}&\mathbb{E}_q[\log \mathrm{Dirichlet}_{V_f}(\beta_{kf\cdot}|A)]\\ &\quad = \log\Gamma(\sum_{v=1}^{V_f}A_v) - \sum_{v=1}^{V_f}\log\Gamma(A_v) + \sum_{v=1}^{V_f}(A_v-1)\left[\psi(\lambda_{kfv}) - \psi(\sum_{u=1}^{V_f}\lambda_{kfu})\right]\\ &\mathbb{E}_q[\mathbb{1}\{z_{dr}=k\}\mathbb{1}\{x_{drf}=v\}\log(\beta_{kfv})]\\ &\quad = \phi_{drk}\mathbb{1}\{x_{drf}=v\}\left[\psi(\lambda_{kfv}) - \psi(\sum_{u=1}^{V_f}\lambda_{kfv})\right]\\ &\mathbb{E}_q[\mathbb{1}\{z_{dr}=k\}\log(\phi_{drk})]\\ &\quad = \phi_{drk}\log(\phi_{drk})\\ &\mathbb{E}_q[\log \mathrm{Dirichlet}_{V_f}(\beta_{kf\cdot}|\lambda_{kf\cdot})]\end{aligned}$$

$$= \log\Gamma(\sum_{v=1}^{V_f}\lambda_{kfv}) - \sum_{v=1}^{V_f}\log\Gamma(\lambda_{kfv}) + \sum_{v=1}^{V_f}(\lambda_{kfv}-1)\left[\psi(\lambda_{kfv}) - \psi(\sum_{u=1}^{V_f}\lambda_{kfu})\right].$$

B.2 Coordinate ascent

We find a local maximum of the ELBO via coordinate ascent in each dimension of the variational parameters: λ, ϕ. This method is sometimes known as *batch* variational inference.

First we look at λ; the partial derivative of the ELBO with respect to λ_{kfv} is

$$\begin{aligned}
\frac{\partial}{\partial\lambda_{kfv}}\text{ELBO}(\phi,\lambda) &= (A_v - 1)\psi_1(\lambda_{kfv}) + \left[\sum_{u=1}^{V_f}(A_u - 1)\right]\psi_1(\sum_{u=1}^{V_f}\lambda_{kfu}) \\
&+ \sum_{d=1}^{D}\sum_{r=1}^{R_d}\phi_{drk}\mathbb{1}\{x_{drf} = v\}\left[\psi_1(\lambda_{kfv}) - \psi_1(\sum_{u=1}^{V_f}\lambda_{kfu})\right] \\
&+ \sum_{u:u\neq v}\sum_{d=1}^{D}\sum_{r=1}^{R_d}\phi_{drk}\mathbb{1}\{x_{drf} = u\}\left[-\psi_1(\sum_{t=1}^{V_f}\lambda_{kft})\right] \\
&- \psi(\sum_{u=1}^{V_f}\lambda_{kfu}) + \psi(\lambda_{kfv}) \\
&- \left\{(\lambda_{kfv} - 1)\cdot\left[\psi_1(\lambda_{kfv}) - \psi_1(\sum_{u=1}^{V_f}\lambda_{kfu})\right]\right\} \\
&- \left[\psi(\lambda_{kfv}) - \psi(\sum_{u=1}^{V_f}\lambda_{kfu})\right] \\
&- \sum_{u:u\neq v}(\lambda_{kfu} - 1)\psi_1(\sum_{t=1}^{V_f}\lambda_{kft}) \\
&= \psi_1(\lambda_{kfv})\left[A_v - \lambda_{kfv} + \sum_{d=1}^{D}\sum_{r=1}^{R_d}\phi_{drk}\mathbb{1}\{x_{drf} = v\}\right] \\
&- \psi_1(\sum_{u=1}^{V_f}\lambda_{kfu})\sum_{u=1}^{V_f}\left[A_u - \lambda_{kfu} + \sum_{d=1}^{D}\sum_{r=1}^{R_d}\phi_{drk}\mathbb{1}\{x_{drf} = u\}\right]
\end{aligned}$$

This quantity will be zero for

$$\lambda_{kfv} \leftarrow A_v + \sum_{d=1}^{D}\sum_{r=1}^{R_d}\phi_{drk}\mathbb{1}\{x_{drf} = v\}.$$

Next we consider ϕ. The partial derivative of the ELBO with respect to ϕ_{drk} is

$$\frac{\partial}{\partial \phi_{drk}} \text{ELBO}(\phi, \lambda) = \sum_{f=1}^{F} \sum_{v=1}^{V_f} \mathbb{1}\{x_{drf} = v\} \left[\psi(\lambda_{kfv}) - \psi(\sum_{u=1}^{V_f} \lambda_{kfu}) \right] - \log(\phi_{drk}) - 1.$$

This quantity will be zero, and the ϕ_{drk} will sum across k to one,[4] if

$$\phi_{drk} \propto_k \exp \left\{ \sum_{f=1}^{F} \sum_{v=1}^{V_f} \mathbb{1}\{x_{drf} = v\} \left[\psi(\lambda_{kfv}) - \psi(\sum_{u=1}^{V_f} \lambda_{kfu}) \right] \right\}.$$

We note that the ϕ_{drk} satisfy the constraint $\phi_{drk} > 0$ for all k and so form a proper probability distribution across k.

[4] This derivation can be completed using the Lagrange method of multipliers.