TRANSFER LEARNING FOR MULTI-CHANNEL HDP-LDA TOPIC MODELS UNDER STRICT DATA SEPARATION CONSTRAINTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Latent Dirichlet Allocation (LDA) models are increasingly used beyond text, including for discrete tabular data in which samples are treated as documents and multinomial variables as observation channels. Such models have found renewed success in the biomedical domain, where datasets are usually small and targeted to a specific health condition. Often, such models can be improved by including data from external sources and transfer learning is the canonical approach in such situations. However, patient data is extremely sensitive and hosted on isolated sandbox environments whose strict governance rules forbid exporting individual-level data to other systems and require manual approval of every data export request. These restrictions make traditional transfer and federated learning methods for LDA models inapplicable. In this paper, we therefore propose new methods for transfer learning for multi-channel LDA models in the presence of strict data-separation constraints and where no iterative communication between data hosting systems is allowed. We tackle this challenge by extracting sufficient statistics from a model fit on the source dataset and using them as frozen, weighted priors when modeling the target dataset, thus requiring a single data transfer between systems. To approximate the topics that would be obtained by combining the datasets, we tune the weight of the frozen source summaries to match their empirical covariance with the nominal covariance they induce in the target model. For target-topic estimation, we introduce exact auxiliary-variable Gibbs updates that infer the transfer strength and the global concentration of a Hierarchical Dirichlet Process (HDP) prior under the asymmetric concentrations induced by the frozen source summaries. Controlled synthetic experiments across a range of sample sizes and domain shifts characterize the methods' behavior, while a case study on the Cleveland heart disease dataset provides a real-world demonstration.

1 INTRODUCTION

Latent Dirichlet Allocation (LDA) models (Blei et al., 2003) have recently been extended to discrete tabular data (Lu et al., 2016), where samples are treated as documents and multinomial variables as observation channels. This extension has sparked renewed interest in healthcare applications (Zhang et al., 2023; Jiang et al., 2023; Donker et al., 2026) as such a Bayesian approach naturally lends itself to modeling complex and overlapping health conditions while being interpretable, robust to missing data, and adept at extracting information from large and small discrete tabular datasets. These challenges are particularly relevant for healthcare and biomedical data science, where data scarcity remains a crucial limitation (Pallikkavaliyaveetil & Chandrasekaran, 2026) while explainability and interpretability are key requirements for practitioners and decision makers (Zhang et al., 2026).

A common approach to the issue of data scarcity is transfer learning, where a general model is learned on large, generic *source* datasets and subsequently refined on smaller, domain-specific *target* datasets. Transfer learning approaches not only promise to improve modeling of the target cohort, but also enable a better understanding of this cohort in the context of the general population, enabling prediction of risk factors and disease progression, and providing deeper understanding of the disease and advancing the development of innovative treatment strategies. Large population biobanks such as the UK Biobank (Bycroft et al., 2018) and FinnGen (Kurki et al., 2023) constitute common starting grounds for building specific models of patients and diseases. However, strict data governance rules in the biomedical setting constitute a critical obstacle for transfer learning, as they typically restrict data analysis to controlled sandbox environments which forbid the export of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

individual-level data and only allow download of aggregate summaries subject to a lengthy manual approval process (UK Biobank, 2026; FinnGen, 2026b). A similar issue exists on the side of the target cohort, where the data is often proprietary, highly confidential, and hosted on private platforms with similarly strict access control rules (DNAnexus, n.d.; Lifebit Biotech Inc, n.d.). Thus, there is a need to transfer knowledge from one cohort to the other without sharing sensitive patient data and while minimizing the number of cohort-level data round-trips between the respective sandboxes. To solve this problem, we propose to extract sufficient statistics from a model fit on the source dataset and inject them into the target model modulated by a strength parameter λ, thus requiring a single data export request from the source sandbox.

Related Work. Introducing weighted prior information has a rich history in Bayesian modeling (Ibrahim & Chen, 2000; Holmes & Walker, 2017; Hobbs et al., 2011; Schmidli et al., 2014; Syring & Martin, 2019), as does incorporating information from multiple datasets via hierarchical structures (Finkel & Manning, 2009); both routes were applied to LDA specifically (Kang et al., 2012; Wood et al., 2017; Jagarlamudi et al., 2012; Wallach et al., 2009). A key difference between these lines of work and our setting is that we do not allow re-evaluating the likelihood of the source data when fitting the model on the target dataset, thus neither the source sufficient statistics nor their weight can be revised to ensure adequate fitting of both datasets. Frozen source summaries were shown to be helpful in the context of *ad hoc* retrieval of natural language documents (Zhai & Lafferty, 2001) and single-cell data modeling (Min et al., 2023), while a software package implementing this approach for traditional LDA models was recently released (Jones, 2026). Exchanging sufficient statistics is common in federated approaches to LDA (Zhao et al., 2021; Jiang et al., 2019; Wang et al., 2020; Jiang et al., 2021), but requires continuous data transmission between the systems hosting the datasets and is made impossible by the data governance rules we consider. Such rules allow the extraction of parameters of the trained models, making transfer learning via neural topic models an attractive option (Akash et al., 2023; Akash & Chang, 2025). However, these methods do not readily provide uncertainty estimates, do not natively handle missing data, are not interpretable, are prone to overfitting (Hoyle et al., 2022), and the topics they discover are usually of low quality (Wu et al., 2024), making them generally unsuitable for the biomedical setting. Therefore, while existing work has studied transfer learning for topic modeling through hierarchical domain adaptation, discounted historical information, neural topic adaptation, and federated inference, these approaches are inapplicable to our setting due to the multi-channel nature of the data, strict restrictions on data exchange between sandboxes, and requirements on model interpretability and explainability.

Our contributions. We study transfer learning for multi-channel LDA topic models when no document-level information may be exchanged between datasets and communication cost between the respective hosting systems is prohibitively high. We propose extracting sufficient statistics from the source model and injecting them into the target model with weight λ, thus requiring a single data export from the source system. Our contributions around this method are as follows:

1. We implement transfer learning for multi-channel LDA models with Hierarchical Dirichlet Process (HDP) priors (Teh et al., 2006) letting the transfer model adapt its topic usage to the target data by reusing topics from the source or activating new ones according to the complexity of the target dataset and its overlap with the source dataset.
2. We derive an exact auxiliary-variable Gibbs update for the global HDP concentration under the asymmetric prior resulting from the injected frozen source summaries, which makes the standard Escobar & West (1995) update invalid.
3. To improve the estimation of target population topics, we derive an exact Gibbs update that automatically infers the strength λ from both the global topic-weight and emission likelihoods, thus avoiding costly manual tuning.
4. To enable the estimation of topics that would be discovered if it were possible to combine the two datasets, we introduce a tuning procedure for λ that calibrates the prior induced by the source statistics in the target model based on their empirical uncertainty.

We evaluate the methods in synthetic settings spanning several domain shifts and dataset sizes, and provide a real-data case study on the Cleveland heart disease dataset. Although motivated by biomedical data analysis, where strict dataset isolation makes the problem especially consequential, our proposed methods apply to LDA models for multi-channel multinomial data generally, including traditional LDA models for text corpora.

2 PRELIMINARIES

LDA topic models (Blei et al., 2003) were generalized to tabular data (Lu et al., 2016) by modeling documents as union of separate, categorical, observable features called *channels*. In the context of healthcare data, patients correspond to documents and biomarkers, i.e., quantitative measurements related to a patient's health status, to channels; continuous observations are typically discretized (Donker et al., 2026). This formulation is particularly suitable for biomedical data, as modeling each patient as a mixture of topics naturally favors the discovery of complex and overlapping disease conditions (Zhang et al., 2023; Jiang et al., 2023; Donker et al., 2026), but also includes ordinary text LDA as the special case of one multinomial channel with observed token counts.

While these models made use of symmetric Dirichlet priors for the document-topic distribution, we posit that a HDP prior (Teh et al., 2006) is more suitable for transfer learning. The HDP formulation is a natural solution to the requirement of manually tuning the number of topics used in the LDA model when fitting on the source dataset *and* when fitting on the target dataset, made especially challenging in a transfer learning setting where the optimal topic number depends on the complexity of each dataset as well as on their semantic overlap. We hereby formally define the multi-channel LDA-HDP topic model using a finite approximation to the HDP structure with K topics.

Definition 2.1 (Multi-channel LDA-HDP topic model with multinomial emissions)**.** *Let D be the number of documents, C the number of channels, V_c the vocabulary size of channel c, and N_{dc} the number of categorical draws observed for channel c in document d. Let $w_{dci} \in \{1, \dots, V_c\}$ be the observed categorical draw $i \in [1, N_{dc}]$ in document $d \in [1, D]$ and channel $c \in [1, C]$. Denote by K the number of topics in the finite symmetric HDP approximation, $\boldsymbol{\beta}_{kc}$ the prior concentrations for the emissions of topic $k \in [1, K]$ in channel $c \in [1, C]$, γ and τ the global and document-level HDP concentration parameters, and $\rho_k = \gamma/K$ the truncated HDP weights. The generative model is the following:*

1. *Sample global topic proportions $\boldsymbol{m} \sim \mathrm{Dir}(\boldsymbol{\rho})$.*
2. *Sample document-topic distribution $\boldsymbol{\pi}_d \mid \boldsymbol{m} \sim \mathrm{Dir}(\tau\, \boldsymbol{m})$ for each document d.*
3. *Sample latent topic $z_{dci} \sim \mathrm{Cat}(\boldsymbol{\pi}_d)$ for each draw i in document d and channel c.*
4. *Sample topic distribution $\boldsymbol{\phi}_{kc} \sim \mathrm{Dir}(\boldsymbol{\beta}_{kc})$ for each topic k and channel c.*
5. *Sample observation $w_{dci} \mid z_{dci} = k \sim \mathrm{Cat}(\boldsymbol{\phi}_{kc})$ for each draw (d, c, i).*

This formulation approaches standard LDA with prior $\boldsymbol{\pi}_d \sim \mathrm{Dir}(\alpha\, \mathbf{1}_K)$ when $\gamma \gg K$ and $\tau = \alpha K$; the original multi-channel construction of Lu et al. (2016) is recovered when $N_{dc} \in \{0, 1\}$, and ordinary bag-of-words LDA (Griffiths & Steyvers, 2004) is the special case $C = 1$ with arbitrary N_{d1}.

In this paper we consider models fit through collapsed Gibbs sampling, an iterative procedure which repeatedly samples each latent variable z_{dci} from its conditional distribution given all others while marginalizing over the topic distributions $\boldsymbol{\pi}_d$ and topic-channel emission parameters $\boldsymbol{\phi}_{kc}$. For the HDP structure, auxiliary variables called *tables* are introduced to bridge $\boldsymbol{z}$ and $\boldsymbol{m}$ in an analogy where documents are restaurants, observations are customers, topics are dishes, customers sitting at a table consume the same dish, and customers are more likely to sit at a table where many customers are already sitting (Teh et al., 2006). A complete description of this algorithm is in Procedure A.1. We assume that J samples are retained to approximate the posterior and use the superscript (j) for the value of a variable after the j-th MCMC sweep when necessary.

Downstream document-level tasks such as clustering and prediction use the posterior mean of $\boldsymbol{\pi}_d$, while model interpretability and explanation uses topic signatures, that is, the posterior mean of $\boldsymbol{\phi}_{kc}$ for each channel, defined as follows.

Definition 2.2 (Topic Signature)**.** *Let $n_{kcv}^{(j)} := \sum_{d=1}^{D} \sum_{i=1}^{N_{dc}} \mathbf{1}[w_{dci} = v \wedge z_{dci}^{(j)} = k]$ be the number of observations of level v in channel c assigned to topic k at MCMC sweep j. The topic signature $\hat{\boldsymbol{\phi}}_{kc}$ of channel c in topic k is defined as the posterior mean of the topic-channel emission parameters averaged over retained MCMC samples:*

$$\hat{\phi}_{kcv} := \frac{1}{J} \sum_{j=1}^{J} \hat{\phi}_{kcv}^{(j)}, \qquad \hat{\phi}_{kcv}^{(j)} := \frac{\beta_{kcv} + n_{kcv}^{(j)}}{\sum_{v'=1}^{V_c} \left(\beta_{kcv'} + n_{kcv'}^{(j)} \right)}. \tag{1}$$

As topic labels are exchangeable, their numeric indices may switch between MCMC sweeps. Therefore, per-sweep signatures $\hat{\boldsymbol{\phi}}_{kc}^{(j)}$ are aligned to the running average of the previously aligned signatures via the Hungarian algorithm before accumulation.

Topic signatures are the main focus of our work: our goal is to improve the signatures learned from one dataset by using information from an external dataset, when the two cannot be combined together to fit a joint model.

3 METHODS

Let $\mathcal{D}^{(s)} = \{w_{dci}^{(s)}\}$ and $\mathcal{D}^{(t)} = \{w_{dci}^{(t)}\}$ denote source and target datasets with $D^{(s)}$ and $D^{(t)}$ documents respectively, and $\boldsymbol{\phi}_{kc}^{(s)}$ and $\boldsymbol{\phi}_{kc}^{(t)}$ the corresponding population-level topic signatures for channel c in topic k. We do not assume topics to be identical or fully overlapping between datasets. We use superscripts (s) and (t) to indicate quantities computed on or related to source or target datasets respectively, combined with the superscript (j) for the MCMC sweep when necessary, but we strive to drop superscripts to simplify notation when it is clear from the context on which dataset we are referring to or the distinction is not necessary.

Our proposed transfer learning procedure starts by fitting a source model on the source dataset $\mathcal{D}^{(s)}$ only. From the retained MCMC samples we extract sufficient statistics that summarize the posterior over the source model's latent variables $\boldsymbol{z}^{(s)}$ aggregated over all documents and draws.

Definition 3.1 (Transfer Prior)**.** *The transfer prior extracted from the source dataset $\mathcal{D}^{(s)}$ is defined as the posterior mean of the source model's sufficient statistics, scaled by a discount factor $\lambda \geq 0$. Each retained source sweep is aligned as described in Definition 2.2 before these sufficient statistics are averaged, so the source counts retain consistent topic labels across sweeps. For the topics:*

$$s_{kcv} = \frac{1}{J}\sum_{j=1}^{J}\sum_{d=1}^{D^{(s)}}\sum_{i=1}^{N_{dc}^{(s)}} \mathbf{1}[w_{dci}^{(s)} = v]\; p(z_{dci}^{(s,j)} = k \mid \textit{rest}^{(s,j)}), \qquad \tilde{\beta}_{kcv} := \beta_{kcv} + \lambda s_{kcv} \tag{2}$$

Where $\textit{rest}^{(s,j)}$ denotes all other variables and assignments held fixed when evaluating the conditional topic probability at source sweep j (detailed in Appendix A). For later use, define the source total for each topic-channel block by $S_{kc} := \sum_{v=1}^{V_c} s_{kcv}$. Let $t_{dk}^{(s,j)}$ be the number of tables in document d serving topic k at sweep j in the source dataset. The corresponding unscaled source summary is

$$r_k := \frac{1}{J}\sum_{j=1}^{J}\sum_{d=1}^{D^{(s)}} t_{dk}^{(s,j)}, \qquad \tilde{\rho}_k := \lambda r_k + \gamma/K \tag{3}$$

The quantities $\tilde{\boldsymbol{\beta}}$ and $\tilde{\boldsymbol{\rho}}$ are then used in place of $\boldsymbol{\beta}$ and $\boldsymbol{\rho}$ in the generative model of Definition 2.1 when it is fit on the target dataset $\mathcal{D}^{(t)}$. In this paper we only consider a single global discount factor λ for the primary reason of being more robust to noisy estimation, however the methods are easily adapted to topic-specific or topic-channel-specific weights. The target model may use a larger K compared to the source model to provide additional unpopulated topics, allowing the HDP structure to freely choose whether to reuse pre-initialized topics from the source dataset, or to introduce entirely new topics that apply exclusively to observations in the target dataset.

Our transfer learning approach can therefore be interpreted as a data-driven prior with λ playing the role of a concentration. This transfer prior contains no document-level information and thus adheres to data export regulations of major biobanks (UK Biobank, 2026; FinnGen, 2026a), and privacy can be further enhanced with suitably analyzed perturbation or statistical disclosure-control mechanisms (Zhao et al., 2021; Hundepool et al., 2026) when required. The introduction of such transfer prior into the target model invalidates some of the traditional sampling machinery for the HDP, an issue which we tackle next.

3.1 UPDATED GLOBAL CONCENTRATION SAMPLING FOR THE HDP

While the model of Definition 2.1 was described with fixed hyperparameters γ and τ, these quantities can also be sampled during inference to adapt topic usage to the complexity of the data distribution (Teh et al., 2006) using the standard update rules proposed by Escobar & West (1995) and included in Procedure A.1.

Under transfer, the symmetric global prior is augmented by the frozen source table summaries of Equation (3), making the finite-truncation prior effectively asymmetric:

$$\boldsymbol{m} \mid \lambda, \gamma \sim \text{Dirichlet}\Big(\lambda r_1 + \frac{\gamma}{K}, \ldots, \lambda r_K + \frac{\gamma}{K}\Big). \tag{4}$$

Crucially, when any transferred table offset λr_k is nonzero the standard Escobar–West update for γ (Proposition A.3) is no longer valid as it assumes that γ scales a fixed normalized global base measure.

We introduce auxiliary variables to enable conjugate Gibbs sampling for the collapsed finite-K conditional of γ. Specifically, we leverage a beta auxiliary for the total-concentration term (Wang et al., 2024), Bernoulli indicators whose sums give auxiliary table counts for each topic (Zhou & Carin, 2015), and binomial draws that attribute those counts to either the transferred mass λr_k or the unknown mass γ/K. Together these steps yield a Gamma update for γ, derived in Proposition B.6.

Proposition 3.1 (Gibbs update for γ)**.** *Let $q_k = \sum_d t_{dk}^{(t)}$ be the number of target tables serving topic k, $Q = \sum_k q_k$, and $R = \sum_k r_k$. Consider the global table model $\gamma \sim \mathrm{Gamma}(a_\gamma, b_\gamma)$ in shape–rate parameterization, $\boldsymbol{m} \mid \gamma, \lambda \sim \mathrm{Dirichlet}(\lambda r_1 + \gamma/K, \ldots, \lambda r_K + \gamma/K)$, and $\boldsymbol{q} \mid Q, \boldsymbol{m} \sim \mathrm{Multinomial}(Q, \boldsymbol{m})$. With $\boldsymbol{m}$ integrated out, for $Q > 0$ draw*

$$u \sim \mathrm{Beta}(Q, \lambda R + \gamma), \tag{5}$$

$$\ell_k = \sum_{i=0}^{q_k - 1} \xi_{ki}, \quad \xi_{ki} \overset{\text{ind}}{\sim} \mathrm{Bernoulli}\left(\frac{\lambda r_k + \gamma/K}{\lambda r_k + \gamma/K + i}\right), \tag{6}$$

$$g_k \sim \mathrm{Binomial}\left(\ell_k, \frac{\gamma/K}{\lambda r_k + \gamma/K}\right), \tag{7}$$

$$\gamma \mid u, \boldsymbol{g}, \boldsymbol{q}, \lambda \sim \mathrm{Gamma}\left(a_\gamma + \sum_k g_k,\ b_\gamma - \log(1-u)\right), \tag{8}$$

for $k = 1, \ldots, K$, taking $\ell_k = g_k = 0$ if $q_k = 0$.

The key decision for transfer learning remains the choice of the transfer strength λ, which controls the influence of the source dataset on the target model. One may expect $\lambda = 1$ to be a sensible choice, however experimental evidence in Section 4 consistently shows that both $\lambda < 1$ and $\lambda > 1$ can result in more faithful topic estimates. Moreover, the best choice of λ depends on whether the practitioner's intention is to improve the estimation of topics in the target dataset, or of the topics that would be obtained by combining the datasets. We now turn to each of these problems.

3.2 Tuning the transfer strength for estimating target topics

When the estimand of interest is the target emissions $\boldsymbol{\phi}^{(t)}$, the direct availability of the target dataset makes it possible to directly evaluate the quality of the fit and consequently to tune the transfer strength λ with traditional methods such as grid search. However, such manual tuning procedures are computationally expensive because they involve fitting multiple models and, moreover, do not offer desirable Bayesian inference properties such as posterior regularization and uncertainty quantification (Gelman et al., 2013). Therefore, we introduce a procedure to sample λ directly alongside the other variables while fitting the model.

Although the collapsed conditional of λ is not itself conjugate, applying similar augmentation schemes as used for Proposition 3.1 yields a Gamma full conditional, the complete derivation of which is given in Appendix B.3.

Proposition 3.2 (Gibbs update for λ)**.** *Consider the model $\lambda \sim \mathrm{Gamma}(a_\lambda, b_\lambda)$ in shape–rate parameterization, with global weights $\boldsymbol{m} \mid \lambda, \gamma \sim \mathrm{Dirichlet}(\lambda r_1 + \gamma/K, \ldots, \lambda r_K + \gamma/K)$ generating the table totals $\boldsymbol{q}$, and emissions $\boldsymbol{\phi}_{kc} \mid \lambda \sim \mathrm{Dirichlet}(\boldsymbol{\beta}_{kc} + \lambda \boldsymbol{s}_{kc})$ generating the emission counts $\boldsymbol{n}_{kc}$, both through multinomial likelihoods. With the global weights and emission distributions integrated out, apply the beta, Bernoulli-table, and binomial-attribution augmentations described above to the global table counts $\boldsymbol{q}$ and separately to each topic–channel emission count block $\boldsymbol{n}_{kc}$. Their contributions add to the shape and rate of a Gamma full conditional for λ. The complete sampling scheme is provided in Proposition B.7.*

The transfer learning method for estimating target topics is then as follows.

Method 3.1 (Transfer learning for estimating target topics)**.** *Target topics are estimated as:*

1. *Fit the LDA-HDP model of Definition 2.1 via Procedure A.1 on the source dataset $\mathcal{D}^{(s)}$*
2. *Compute the transfer prior in Definition 3.1 with initial value $\lambda = a_\lambda / b_\lambda$.*
3. *Fit a LDA-HDP model on the target dataset $\mathcal{D}^{(t)}$ using the transfer prior computed in Step 2 and the Gibbs updates in Propositions 3.1 and 3.2 (listing in Procedure B.2).*

This approach is enabled by the fact that the estimated target topics can be directly evaluated using the target dataset. When the goal is to estimate the joint topics, that is, the topics that would be discovered if we were able to pool source and target datasets together, the target dataset alone is insufficient to evaluate the quality of the fit and the procedure outlined above cannot therefore be expected to achieve faithful estimates.

3.3 Tuning the transfer strength for estimating joint topics

The setting we consider forbids re-evaluating the source likelihood in light of new data, invalidating most existing approaches. The source sufficient statistics must stay frozen and the only lever we have for adjusting the influence of the source data is the transfer strength λ. Note how injecting the transfer prior $\tilde{\boldsymbol{\beta}}_{kc}$ into the transfer model of Definition 2.1 makes the prior for $\boldsymbol{\phi}_{kc}^{(t)}$ a Dirichlet with concentration $\boldsymbol{\beta}_{kc} + \lambda \boldsymbol{s}_{kc}$. This centers the prior on the posterior mean of the emissions estimated by the source model, which is the desired effect, however it also fixes the covariance of this distribution to

$$\Sigma_{kc}(\lambda) = \frac{A_{kc}(\lambda)\,\mathrm{diag}(\boldsymbol{a}_{kc}(\lambda)) - \boldsymbol{a}_{kc}(\lambda)\boldsymbol{a}_{kc}(\lambda)^{\top}}{A_{kc}(\lambda)^2\,[A_{kc}(\lambda)+1]} \tag{9}$$

where $a_{kcv}(\lambda) = \beta_{kcv} + \lambda s_{kcv}$ and $A_{kc}(\lambda) = \sum_v a_{kcv}(\lambda)$. As the source counts $\boldsymbol{s}_{kc}$ increase, the prior for $\boldsymbol{\phi}_{kc}^{(t)}$ inevitably becomes more and more concentrated, even though the uncertainty of the source model may not support such a narrow concentration.

To avoid this effect and give due weight to the source counts, we choose λ such that the nominal covariance in Equation (9) matches the observed posterior covariance of the estimated source topic signatures. This covariance shall be computed from additional posterior draws during sampling of the source model to include uncertainty both in the latent allocations $\boldsymbol{z}^{(s,j)}$ and in the emissions $\hat{\boldsymbol{\phi}}^{(s,j)}$ conditional on those allocations. At each retained MCMC sweep j of the source model, let

$$\tilde{\boldsymbol{\phi}}_{kc}^{(s,j)} \sim \mathrm{Dir}(\boldsymbol{\beta}_{kc} + \boldsymbol{n}_{kc}^{(s,j)}) \tag{10}$$

be a draw from the marginal source posterior ; then, their empirical covariance is computed as

$$\widehat{\Sigma}_{kc} := \frac{1}{J-1}\sum_{j=1}^{J}\left(\tilde{\boldsymbol{\phi}}_{kc}^{(s,j)} - \bar{\boldsymbol{\phi}}_{kc}^{(s)}\right)\left(\tilde{\boldsymbol{\phi}}_{kc}^{(s,j)} - \bar{\boldsymbol{\phi}}_{kc}^{(s)}\right)^{\top}, \qquad \bar{\boldsymbol{\phi}}_{kc}^{(s)} := \frac{1}{J}\sum_{j=1}^{J}\tilde{\boldsymbol{\phi}}_{kc}^{(s,j)} \tag{11}$$

The counts and emission draws in these expressions use the same per-sweep alignment to the running average described in Definition 2.2, thus the empirical covariance measures posterior variation for a consistent labeling of source topic rather than variation caused by changing topic indices. We then propose a moment-matching rule that chooses λ to match these two covariances.

Proposition 3.3 (Dirichlet covariance matching rule)**.** *Let $\widehat{\Sigma}_{kc}$ be the empirical covariance of a topic-channel block (k, c) estimated via Equation* (11) *and $\Sigma_{kc}(\lambda)$ the nominal Dirichlet covariance of Equation* (9) *induced by a transfer prior with strength λ; then, a robust global covariance-matching transfer strength is*

$$\hat{\lambda} = \arg\min_{\lambda>0} \mathrm{Median}_{(kc):S_{kc}>0}\ \frac{1}{V_c}\left\|\widehat{\Sigma}_{kc} - \Sigma_{kc}(\lambda)\right\|_F^2 \tag{12}$$

where $\|\cdot\|_F$ is the Frobenius norm of a matrix and V_c the number of levels of channel c.

We employ median aggregation over all topics and channels to provide robustness towards poorly-estimated covariances of individual topic-channel blocks. [1] Note that $\hat{\lambda}$ is computed entirely within the source dataset sandbox and makes no assumptions regarding the target dataset: its only purpose is to calibrate the induced prior dispersion to match its observed uncertainty. The complete method for estimating joint topics is therefore:

Method 3.2 (Transfer learning for estimating joint topics)**.** *Joint topics are estimated as follows:*

1. *Fit the LDA-HDP model of Definition 2.1 via Procedure A.1 on the source dataset $\mathcal{D}^{(s)}$*
2. *Estimate the global transfer strength $\hat{\lambda}$ via the Dirichlet covariance matching heuristic in Proposition 3.3*
3. *Compute the transfer prior in Definition 3.1 with the value of $\hat{\lambda}$ computed in Step 2.*
4. *Fit a LDA-HDP model on the target dataset $\mathcal{D}^{(t)}$ using the transfer prior computed in Step 3 and the update rule in Proposition 3.1 for γ (listing in Procedure B.1)*

[1] Preliminary experiments using weighted average pooling with optional trimming achieved similar performance, therefore we recommend the simplest option.

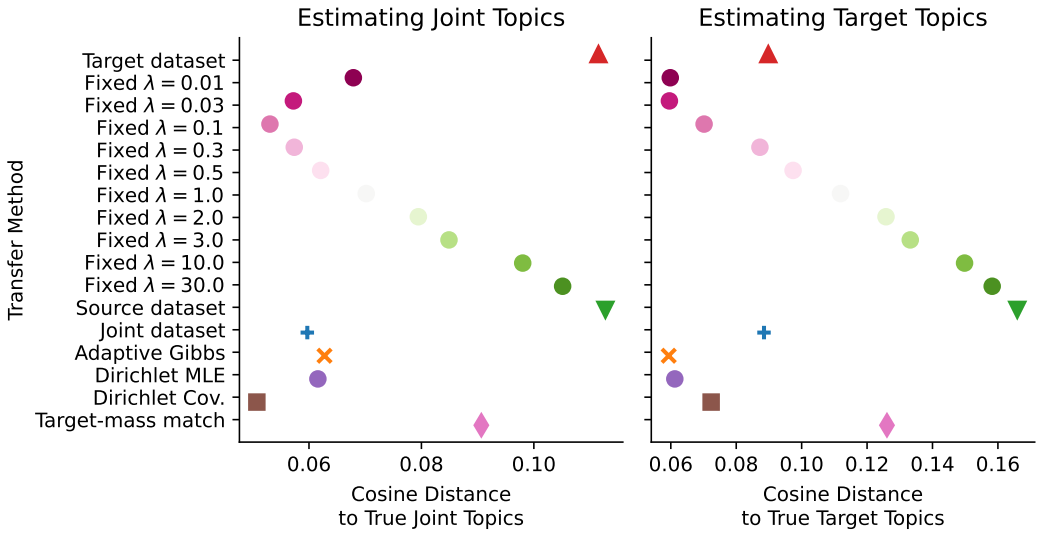


Figure 1: Average cosine distance (x-axis) to true population topics when estimating joint (left) and target (right) topics for different transfer learning methods (y-axis) for the synthetic datasets. The methods *Source dataset* and *Target dataset* are the baselines without transfer learning, and *Joint dataset* are models fit on the combined dataset. *Adaptive Gibbs* is Method 3.1, and *Dirichlet Cov.* is Method 3.2. The *Target-mass match* baseline sets λ to the value matching the total scaled source count mass to the target token counts, and *Dirichlet MLE* maximizes the median block log-density of the source posterior draws.

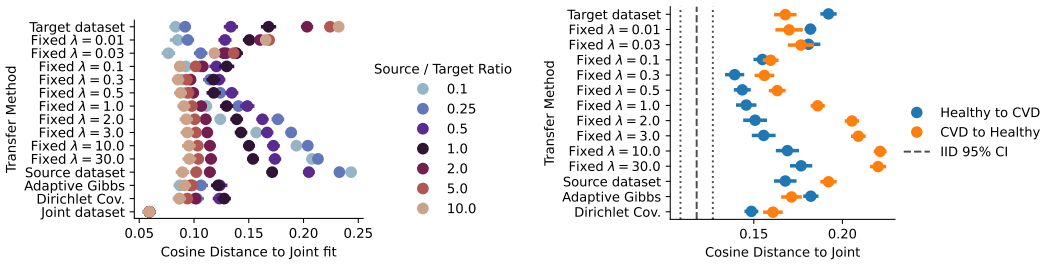
4 Experimental validation

We start with validating our proposed methods on synthetic datasets, where knowing the true population emissions allows us to directly compare the methods by quality of discovered topics. We then present a case study on the Cleveland heart disease dataset (Carrion, 2022), comparing transfer results with models fit on the combined data. We construct an oracle via grid search over λ (Appendix C.1), and note that this is not available in practice when the joint data cannot be accessed. Models fit only on either the source or target dataset serve as baselines without transfer. Besides oracle and baselines, for each pair we fit the Adaptive Gibbs procedure of Method 3.1, and the covariance heuristics of Method 3.2. We used the same number of topics K between source and target datasets, and compared topic signatures between model fits via the cosine distance after concatenating the emissions of all channels into a single vector for each topic and aligning these via the Hungarian algorithm. Model fitting details in Appendix C.1.

4.1 Synthetic datasets

We sampled synthetic datasets following a Hierarchical Dirichlet generative process in which source and target population topics are sampled conditioned on shared global topics (details in Appendix C.3). We varied the concentrations of all Dirichlet distributions over a grid, independently between source and target datasets. This resulted in a variety of different scenarios, including small and large domain shift, peaked and well-separated to flat and overlapping topics, and few to many topics used for each document. We chose the size of the datasets to compare two different regimes for each dataset: 100 documents did not allow for a precise estimation of population emissions, while 1,000 documents did. We can thus understand how transfer learning behaves as topics in the source, target, both, or neither are properly estimated. We considered 972 combinations of dataset generation parameters sampling 6 pairs of source and target datasets for each, giving 5,832 dataset pairs in total. We also tested an alternative to the covariance matching, *Dirichlet MLE*, that set λ by maximizing the log-density of the posterior draws in Equation (10) under $\mathrm{Dir}(\boldsymbol{\beta} + \lambda \boldsymbol{s})$ (Appendix C.2). Finally, we introduced a target-mass matching baseline that set λ to the ratio of total observed target to source tokens, making the weight of the transferred source counts equal to that of the target counts (Appendix C.2).

Estimating joint emissions. When estimating joint topics, $\lambda = 0.1$ performed best on average (Figure 1), however when the source dataset was much larger than the target the best strength in-



(a) Distance to joint fit for IID setting.

(b) Distance to joint fit for OOD setting.

Figure 2: Results on the Cleveland heart disease dataset: Cosine distance (x-axis) between topic signatures discovered on the full dataset and those discovered by transfer learning for different transfer methods (y-axis). a Result in the IID setting with different relative sizes of source and target datasets (hue colors). Error bars denote standard error of the mean. b Result in the OOD setting when transferring from healthy to CVD (blue) or CVD to healthy (orange). The dashed black line is the distance of an IID fit with comparable source/target dataset sizes, and the dotted lines indicate its 95% confidence interval.

creased to 1-4 (Table C.1). The best grid search results achieved average cosine distances to the true topics of 0.047 compared to 0.060 for the joint fit, while the Dirichlet covariance heuristic achieved average distance of 0.051, lower than any fixed value of λ. The Dirichlet maximum-likelihood baseline achieved an average distance of 0.062, while the source-only and target-only baselines were at average distances of 0.113 and 0.111 respectively. [2] Disaggregated results (Figure C.1) show that when the source dataset was much larger than the target dataset, any value of λ greater than 0.3-0.5 performed comparably well regardless of the domain shift, and all forms of transfer improved over target-only fits. Moreover, joint fits underperformed compared to transfer when the domain shift was large. The Adaptive Gibbs procedure performed on average surprisingly well with a distance of 0.063, but disaggregated results show that it particularly struggled when the domain shift was large (Table C.2) or the target dataset was small (Table C.3).

Estimating target emissions. Grid search obtained again the best performance for estimating the target topics with average distance of 0.054, and was closely followed by the Adaptive Gibbs procedure whose average performance was 0.060. Both transfer methods outperformed the target-only baseline, which achieved an average distance of 0.090 to the true topics, and the covariance matching rule obtained distances of 0.072 on average. The Dirichlet MLE baseline (0.061) and small fixed strengths ($\lambda \leq 0.03$) performed comparably to Adaptive Gibbs on average. On average, joint and target-only fits performed similarly with average distances of 0.088 and 0.090, however disaggregated results (Figure C.2) show that joint fits performed better when the source dataset was large and the domain shift small and worse in the opposite scenario.

4.2 Cleveland Case Study

As a real-data case study, we used the comprehensive cohort of the Cleveland heart disease dataset containing 11 biomarkers for 1,190 patients (Janosi et al., 1988; Detrano et al., 1989; Carrion, 2022) (preprocessing information in Appendix C.4). As the true population topic signatures are unknown, we compared the topic signatures discovered by transfer models with those discovered by models fit on the full dataset. This changes the interpretation of evaluation metrics from distance to *the true population topics* to distance to *the information available in the data* which can be extracted from a comparable model. This is particularly relevant for estimation of target topics, as experiments on synthetic data showed that the target-only fit is not necessarily the best estimate of the true topic signatures of the target population. We note that none of the existing methods for LDA transfer learning (Wood et al., 2017; Jones, 2026; Min et al., 2023; Kang et al., 2012) is applicable to such multi-channel data. We artificially divided the full dataset following two protocols emulating in-distribution and out-of-distribution transfer learning, and studied the effect of HDP hyperparameter optimization by comparing sampled hyperparameters with results achieved when fixing them to the posterior averages found on whole-dataset fits.

[2] Because of the large number of experiments, most differences are statistically significant even when practically irrelevant. Standard errors reported in the appendix.

IID scenario. We partitioned the dataset into source and target completely at random such that the relative sizes of the two varied from 10% to 90%, and for each we repeated the experiment with three different random splits, and three chains with different random initialization for each split.

Across all relative sizes, the best grid search results had an average cosine distance of 0.098 to the joint fits, while the covariance matching and Adaptive Gibbs methods had an average cosine distance of 0.103 and 0.104 respectively ($p = 0.657$, Figure 2a). This can be explained by noting that with such an IID splitting protocol, target topics and joint topics are effectively the same, and thus both procedures are estimating the same quantity. Incidentally, this also supports the validity of the covariance matching heuristic. The target-only and source-only dataset baselines approached the performance of transfer learning as the respective dataset size increased to 90% of the total (Table C.4). The posterior HDP hyperparameters on the target dataset were very similar across all transfer methods when the source dataset was large, whereas with large target datasets, low values of λ favored more global topics and few topics per document, with the trend reversing as λ was larger (Figure C.3). Compared to fixed values, sampled HDP hyperparameters had a small but significant negative performance hit both for the grid search oracle and for the covariance matching heuristic ($p = 0.006$ and $p = 0.018$, paired permutation test, Figure C.4). The distribution of global topic weights $\boldsymbol{m}$ was similar between source and transfer fits, and was only affected by the relative dataset sizes rather than the value of λ (Figures C.5 and C.6).

OOD Scenario. To simulate transfer learning between datasets with significant domain shift, we partitioned the full dataset according to the target label into healthy patients ($N = 629$, 53%) and those with Cardiovascular diseases (CVD, $N = 561$, 47%). Both of these groups served as source and target datasets for transfer learning, and as before we evaluated the quality of the topic signatures discovered by the transfer model with respect to those discovered by models fit on the full dataset, fitting three chains for each model.

Transferring from healthy to CVD and vice-versa, grid search reached distances to the joint fits of 0.139 and 0.156 and covariance matching 0.149 and 0.161, and in both cases the difference was not statistically significant ($p = 0.094$ and $p = 0.452$), while source- and target-only baselines performed worst at 0.168-0.192 (Figure 2b, Table C.5). Grid search on a comparable IID split reached a distance of 0.118. The difference in performance could be largely explained by HDP hyperparameter optimization, as with fixed hyperparameters there was no statistically significant difference between grid search on IID and OOD splits, although Dirichlet covariance matching was slightly worse when transferring from healthy to CVD patients ($p = 0.013$, Figure C.7). The direction of transfer affected global topic weights $\boldsymbol{m}$: when transferring to CVD patients, both the Adaptive Gibbs model and the fixed $\lambda = 0.3$ methods preferred to re-use topics discovered in healthy patients, while in the other direction both models preferred using more topics (Figures C.8 and C.9).

5 Conclusion

We focused on multi-channel LDA topic models with HDP prior and studied the issue of fitting models on two separate datasets that cannot be combined. We proposed to initialize the target model with appropriately scaled sufficient statistics extracted from the source model, and adapted HDP hyperparameter sampling accordingly. We then proposed two ways of automatically setting the right transfer strength depending on whether the final goal is to improve estimation of the topics in the target dataset, or estimate the topics that would be obtained if source and target datasets could be combined.

On synthetic data, tuned transfer consistently improved topic estimation over single-dataset baselines, with gains depending on dataset sizes and domain shift. When estimating joint topics, covariance matching came close to the grid search oracle and both outperformed fits on the combined data, especially under large domain shift; when estimating target topics, the Adaptive Gibbs procedure was close to the oracle and clearly better than both target-only and combined fits. On the Cleveland dataset, both procedures approached the oracle in the IID setting, and covariance matching was statistically indistinguishable from it in the OOD setting, where the remaining gap to IID transfer was largely attributable to HDP hyperparameter optimization.

In conclusion, our solutions, originally motivated by data governance rules protecting the privacy of patients who donated their data to large population biobanks (Bycroft et al., 2018; Kurki et al., 2023), enable practitioners to fit a multi-channel LDA topic model on one dataset and refine its topics on a separate dataset with a single export of aggregate statistics between the respective hosting systems. Future work focusing on comprehensive benchmarking and model interpretation is needed.

DECLARATIONS

AI Disclosure In this work, we used generative AI tools for proposing or refining hypotheses, formulating mathematical claims, assisting in the writing of proofs, providing critical ingredients for proving mathematical claims, designing or providing feedback on research methodology or experiments, and implementing methods. We have not used generative AI tools for generating synthetic data sets, helping develop theoretical models or conceptual frameworks, cleaning and reformatting dataset, and interpreting results, while assisting with translation, and supporting qualitative and thematic data analysis are not applicable to this work. Additionally, we used generative AI tools for brainstorming, sourcing/searching for information, identifying relevant literature, summarizing or analyzing existing literature, drafting parts of a research paper, editing a research paper to improve readability, creating or editing software code. We have reviewed all AI-assisted work by manually reviewing the literature for related ideas, manually checking all derivations for proofs and the method implementation, manually writing code for running the experiments and analyzing the results, and writing significant parts of the manuscript. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics Statement This work is intended to support ethical biomedical research while protecting the privacy, confidentiality, and autonomy of individuals whose data may contribute to such research. Modeling results are not a substitute for expert clinical judgment and should be independently validated by qualified professionals before being used in research or any decision that could affect an individual. Users are responsible for applying these methods lawfully, responsibly, and in ways that do not cause harm.

Reproducibility Statement The main text specifies the model, transfer-learning procedures, and experimental protocols; complete collapsed Gibbs-sampling updates and derivations are provided in Appendix A, Appendix B.2, and Appendix B.3. The appendix also documents the model-fitting settings, synthetic-data generation, and Cleveland dataset preprocessing in Appendix C.1, Appendix C.3, and Appendix C.4, respectively, while additional experimental results are reported in Appendix C.5 and Appendix C.6. An anonymous implementation of the models, inference procedures, and experimental pipelines is released for peer review only at `https://anonymous.4open.science/r/tabular_topic_modeling_transfer-C3EC` and will be published upon acceptance.

REFERENCES

Pritom Saha Akash and Kevin Chen-Chuan Chang. Understanding cross-domain adaptation in low-resource topic modeling. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5988–6001, 2025.

Pritom Saha Akash, Trisha Das, and Kevin Chen-Chuan Chang. Topicadapt- an inter-corpora topics adaptation approach, 2023. URL https://arxiv.org/abs/2310.04978.

Charles E. Antoniak. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics*, 2:1152–1174, 1974.

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

Clare Bycroft, Colin Freeman, Desislava Petkova, Gavin Band, Lloyd T. Elliott, Kevin Sharp, Allan Motyer, Damjan Vukcevic, Olivier Delaneau, Jared O'Connell, et al. The uk biobank resource with deep phenotyping and genomic data. *Nature*, 562(7726):203–209, 2018. doi: 10.1038/s41586-018-0579-z.

Dustin Carrion. Heart-disease-dataset-(comprehensive), 2022. URL https://www.openml.org/d/43672. OpenML dataset ID 43672, accessed 2026-08-27.

Robert Detrano, Andras Jánosi, William Steinbrunn, Matthias Pfisterer, Johann Schmid, Sarbjit Sandhu, Kimberly Guppy, Stella Lee, and Victor Froelicher. International application of a new probability algorithm for the diagnosis of coronary artery disease. *The American Journal of Cardiology*, 1989.

DNAnexus. Security. compliance. quality., n.d. URL https://www.dnanexus.com/platform-security-compliance. Accessed 2026-08-31.

Hylke C Donker, Vartika Bisht, Om Prakash Dwivedi, Stefanie Mueller, Dorien Neijzen, Zhihao Ding, and Gerton Lunter. Cardiovascular-kidney-metabolic syndrome: candidate subtypes and genetic risk factors. *BMC Medical Genomics*, 2026.

Michael D. Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90(430):577–588, 1995.

Jenny Rose Finkel and Christopher D Manning. Hierarchical bayesian domain adaptation. In *Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*, pp. 602–610, 2009.

FinnGen. Green data users, 2026a. URL https://docs.finngen.fi/where-to-begin../quick-guides/green-data-users. Accessed 2026-08-28.

FinnGen. Red data users, 2026b. URL https://docs.finngen.fi/where-to-begin../quick-guides/red-data-users. Accessed 2026-08-28.

Andrew Gelman, John B. Carlin, Hal S. Stern, David B. Dunson, Aki Vehtari, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, 3 edition, 2013.

Thomas L. Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences*, 101(suppl_1):5228–5235, 2004.

Brian P. Hobbs, Bradley P. Carlin, Sumithra J. Mandrekar, and Daniel J. Sargent. Hierarchical commensurate and power prior models for adaptive incorporation of historical information in clinical trials. *Biometrics*, 67(3):1047–1056, 2011.

Chris C Holmes and Stephen G Walker. Assigning a value to a power likelihood in a general bayesian model. *Biometrika*, 104(2):497–503, 2017.

Alexander Miserlis Hoyle, Pranav Goel, Rupak Sarkar, and Philip Resnik. Are neural topic models broken? In *Findings of the association for computational linguistics: EMNLP 2022*, pp. 5321–5344, 2022.

Anco Hundepool, Josep Domingo-Ferrer, Luisa Franconi, Sarah Giessing, Rainer Lenz, Jane Naylor, Eric Schulte Nordholt, Giovanni Seri, Peter-Paul De Wolf, Reinhard Tent, Andrzej Młodak, Johannes Gussenbauer, and Kamil Wilak. Handbook on statistical disclosure control, 06 2026. URL https://sdctools.github.io/HandbookSDC/.

Joseph G. Ibrahim and Ming-Hui Chen. Power prior distributions for regression models. *Statistical Science*, 15(1):46–60, 2000.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 204–213, 2012.

Andras Janosi, William Steinbrunn, Matthias Pfisterer, and Robert Detrano. Heart disease. UCI Machine Learning Repository, 1988.

Di Jiang, Yuanfeng Song, Yongxin Tong, Xueyang Wu, Weiwei Zhao, Qian Xu, and Qiang Yang. Federated topic modeling. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, pp. 1071–1080. ACM, November 2019. doi: 10.1145/3357384.3357909. URL http://dx.doi.org/10.1145/3357384.3357909.

Di Jiang, Yongxin Tong, Yuanfeng Song, Xueyang Wu, Weiwei Zhao, Qian Xu, and Qiang Yang. Industrial federated topic modeling. *ACM Transactions on Intelligent Systems and Technology*, 12(1):1–22, 2021.

Xilin Jiang, Martin Jinye Zhang, Yidong Zhang, Arun Durvasula, Michael Inouye, Chris Holmes, Alkes L. Price, and Gil McVean. Age-dependent topic modeling of comorbidities in uk biobank identifies disease subtypes with differential genetic risk. *Nature Genetics*, 55(11):1854–1865, October 2023. ISSN 1546-1718. doi: 10.1038/s41588-023-01522-8. URL http://dx.doi.org/10.1038/s41588-023-01522-8.

Tommy Jones. Transfer learning with lda (tlda), 2026. URL https://cran.r-project.org/web/packages/tidylda/vignettes/tLDA.html.

Jeon-Hyung Kang, Jun Ma, and Yan Liu. Transfer Topic Modeling with Ease and Scalability. In *2012 SIAM International Conference on Data Mining (SDM12)*, pp. 564–575, 2012. doi: 10.1137/1.9781611972825.49.

Mitja I. Kurki, Juha Karjalainen, Priit Palta, Timo P. Sipilä, Kati Kristiansson, Kati M. Donner, Mary P. Reeve, Hannele Laivuori, Mervi Aavikko, Mari A. Kaunisto, Anu Loukola, Elisa Lahtela, Hannele Mattsson, Päivi Laiho, Pietro Della Briotta Parolo, Arto A. Lehisto, Masahiro Kanai, Nina Mars, Joel Rämö, Tuomo Kiiskinen, Henrike O. Heyne, Kumar Veerapen, Sina Rüeger, Susanna Lemmelä, Wei Zhou, Sanni Ruotsalainen, Kalle Pärn, Tero Hiekkalinna, Sami Koskelainen, Teemu Paajanen, Vincent Llorens, Javier Gracia-Tabuenca, Harri Siirtola, Kadri Reis, Abdelrahman G. Elnahas, Benjamin Sun, Christopher N. Foley, Katriina Aalto-Setälä, Kaur Alasoo, Mikko Arvas, Kirsi Auro, Shameek Biswas, Argyro Bizaki-Vallaskangas, Olli Carpen, Chia-Yen Chen, Oluwaseun A. Dada, Zhihao Ding, Margaret G. Ehm, Kari Eklund, Martti Färkkilä, Hilary Finucane, Andrea Ganna, Awaisa Ghazal, Robert R. Graham, Eric M. Green, Antti Hakanen, Marco Hautalahti, Åsa K. Hedman, Mikko Hiltunen, Reetta Hinttala, Iiris Hovatta, Xinli Hu, Adriana Huertas-Vazquez, Laura Huilaja, Julie Hunkapiller, Howard Jacob, Jan-Nygaard Jensen, Heikki Joensuu, Sally John, Valtteri Julkunen, Marc Jung, Juhani Junttila, Kai Kaarniranta, Mika Kähönen, Risto Kajanne, Lila Kallio, Reetta Kälviäinen, Jaakko Kaprio, Nurlan Kerimov, Johannes Kettunen, Elina Kilpeläinen, Terhi Kilpi, Katherine Klinger, Veli-Matti Kosma, Teijo Kuopio, Venla Kurra, Triin Laisk, Jari Laukkanen, Nathan Lawless, Aoxing Liu, Simonne Longerich, Reedik Mägi, Johanna Mäkelä, Antti Mäkitie, Anders Malarstig, Arto Mannermaa, Joseph Maranville, Athena Matakidou, Tuomo Meretoja, Sahar V. Mozaffari, Mari E. K. Niemi, Marianna Niemi, Teemu Niiranen, Christopher J. O´Donnell, Ma´en Obeidat, George Okafo, Hanna M. Ollila, Antti Palomäki, Tuula Palotie, Jukka Partanen, Dirk S. Paul, Margit Pelkonen, Rion K. Pendergrass, Slavé Petrovski, Anne Pitkäranta, Adam Platt, David Pulford, Eero Punkka, Pirkko Pussinen, Neha Raghavan, Fedik Rahimov, Deepak Rajpal, Nicole A. Renaud, Bridget Riley-Gillis, Rodosthenis Rodosthenous, Elmo Saarentaus, Aino Salminen, Eveliina Salminen, Veikko Salomaa, Johanna Schleutker, Raisa Serpi, Huei-yi Shen, Richard Siegel, Kaisa Silander, Sanna Siltanen, Sirpa Soini, Hilkka Soininen, Jae Hoon Sul,

Ioanna Tachmazidou, Kaisa Tasanen, Pentti Tienari, Sanna Toppila-Salmi, Taru Tukiainen, Tiinamaija Tuomi, Joni A. Turunen, Jacob C. Ulirsch, Felix Vaura, Petri Virolainen, Jeffrey Waring, Dawn Waterworth, Robert Yang, Mari Nelis, Anu Reigo, Andres Metspalu, Lili Milani, Tõnu Esko, Caroline Fox, Aki S. Havulinna, Markus Perola, Samuli Ripatti, Anu Jalanko, Tarja Laitinen, Tomi P. Mäkelä, Robert Plenge, Mark McCarthy, Heiko Runz, Mark J. Daly, and Aarno Palotie. Finngen provides genetic insights from a well-phenotyped isolated population. *Nature*, 613(7944):508–518, January 2023. ISSN 1476-4687. doi: 10.1038/s41586-022-05473-8. URL http://dx.doi.org/10.1038/s41586-022-05473-8.

Lifebit Biotech Inc. Lifebit trust center, n.d. URL https://lifebit.ai/trust-center/. Accessed 2026-08-31.

Hsin-Min Lu, Chih-Ping Wei, and Fei-Yuan Hsiao. Modeling healthcare data using multiple-channel latent dirichlet allocation. *Journal of Biomedical Informatics*, 60:210–223, April 2016. ISSN 1532-0464. doi: 10.1016/j.jbi.2016.02.003. URL http://dx.doi.org/10.1016/j.jbi.2016.02.003.

Alan Min, Timothy Durham, Louis Gevirtzman, and William Stafford Noble. Matrix prior for data transfer between single cell data types in latent dirichlet allocation. *PLOS Computational Biology*, 19(5):e1011049, May 2023. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1011049. URL http://dx.doi.org/10.1371/journal.pcbi.1011049.

Nazreen Pallikkavaliyaveetil and Sriram Chandrasekaran. Small data, big challenges: Machine- and deep-learning strategies for data-limited drug discovery. *Advanced Drug Delivery Reviews*, 229:115762, February 2026. ISSN 0169-409X. doi: 10.1016/j.addr.2025.115762. URL http://dx.doi.org/10.1016/j.addr.2025.115762.

Heinz Schmidli, Sandro Gsteiger, Satrajit Roychoudhury, Anthony O'Hagan, David Spiegelhalter, and Beat Neuenschwander. Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics*, 70(4):1023–1032, October 2014. ISSN 1541-0420. doi: 10.1111/biom.12242. URL http://dx.doi.org/10.1111/biom.12242.

Nicholas Syring and Ryan Martin. Calibrating general posterior credible regions. *Biometrika*, 106 (2):479–486, 2019.

Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

UK Biobank. Guidance on data downloads and exemptions, 2026. URL https://community.ukbiobank.ac.uk/hc/en-gb/articles/31972311370013-Guidance-on-Data-Downloads-and-Exemptions. Accessed 2026-08-28.

David A. van Dyk and Taeyoung Park. Partially collapsed Gibbs samplers: Theory and methods. *Journal of the American Statistical Association*, 103(482):790–796, 2008.

Hanna M. Wallach, David Mimno, and Andrew McCallum. Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems*, volume 22, pp. 1973–1981, 2009.

Jiahao Wang, Sikun Yang, Heinz Koeppl, Xiuzhen Cheng, Pengfei Hu, and Guoming Zhang. A poisson-gamma dynamic factor model with time-varying transition dynamics, 2024. URL https://arxiv.org/abs/2402.16297.

Yansheng Wang, Yongxin Tong, and Dingyuan Shi. Federated latent dirichlet allocation: A local differential privacy based framework. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):6283–6290, April 2020. ISSN 2159-5399. doi: 10.1609/aaai.v34i04.6096. URL http://dx.doi.org/10.1609/aaai.v34i04.6096.

Justin Wood, Patrick Tan, Wei Wang, and Corey Arnold. Source-lda: Enhancing probabilistic topic models using prior knowledge sources. In *2017 IEEE 33rd international conference on data engineering (ICDE)*, pp. 411–422. IEEE, 2017.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. A survey on neural topic models: methods, applications, and challenges: X. wu et al. *Artificial Intelligence Review*, 57(2):18, 2024.

Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 334–342, 2001.

Kai Zhang, Dongqi Wang, Fuxin Lin, Jue Xie, and Weihua Zhou. A comprehensive review of explainable artificial intelligence in healthcare methods, evaluation, and clinical integration. *iScience*, 29(3):115026, March 2026. ISSN 2589-0042. doi: 10.1016/j.isci.2026.115026. URL http://dx.doi.org/10.1016/j.isci.2026.115026.

Yidong Zhang, Xilin Jiang, Alexander J. Mentzer, Gil McVean, and Gerton Lunter. Topic modeling identifies novel genetic loci associated with multimorbidities in uk biobank. *Cell Genomics*, 3 (8):100371, August 2023. ISSN 2666-979X. doi: 10.1016/j.xgen.2023.100371. URL http://dx.doi.org/10.1016/j.xgen.2023.100371.

Fangyuan Zhao, Xuebin Ren, Shusen Yang, Qing Han, Peng Zhao, and Xinyu Yang. Latent Dirichlet allocation model training with differential privacy. *IEEE Transactions on Information Forensics and Security*, 16:1290–1305, 2021. doi: 10.1109/TIFS.2020.3032021.

Mingyuan Zhou and Lawrence Carin. Negative binomial process count and mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(2):307–320, 2015. doi: 10.1109/TPAMI.2013.211.

APPENDIX

A LDA-HDP GIBBS SAMPLING

All our models are fitted with collapsed Gibbs sampling. This method was introduced for LDA by Griffiths & Steyvers (2004) and adapted for the biomedical setting by Lu et al. (2016), to which we additionally introduced a Hierarchical Dirichlet Process prior (Teh et al., 2006). The sampler is otherwise standard, and we report the full inference procedure here for completeness.

We integrate out the topic–channel emission distributions $\boldsymbol{\phi}_{kc}$ and the document-level topic proportions $\boldsymbol{\pi}_d$. The sampler therefore stores topic assignments and additionally table counts and concentration parameters when the selected model has a DP prior. The collapsed representation is useful here because the transfer counts enter the same predictive distributions as ordinary Dirichlet pseudo-counts.

Definition A.1 (Collapsed sampler state)**.** *For draw (d, c, i), let $z_{dci} \in \{1, \dots, K\}$ be its topic assignment and let*

$$n_{dk} = \sum_{c} \sum_{i=1}^{N_{dc}} \mathbf{1}[z_{dci} = k], \qquad n_{kcv} = \sum_{d} \sum_{i=1}^{N_{dc}} \mathbf{1}[z_{dci} = k, w_{dci} = v] \tag{A.1}$$

denote document-topic and topic-channel-level counts. Superscript $-d, c, i$ means that only draw (d, c, i) is omitted. Missing observations are not counted in n_{kcv}. In an HDP, t_{dk} denotes the number of Chinese restaurant franchise tables in document d serving topic k.

These counts are the sufficient statistics needed by the collapsed updates. In particular, the emission counts are aggregated over documents, so no individual source observation is needed after the transfer prior has been constructed.

A.1 STANDARD LDA GIBBS SAMPLING

For standard LDA, the document-topic proportions have prior $\boldsymbol{\pi}_d \sim \mathrm{Dir}(\alpha_1, \dots, \alpha_K)$ and are integrated out together with the emissions.

Proposition A.1 (Collapsed categorical assignment)**.** *Let β_{kcv} be the base emission prior concentration. For LDA, the conditional distribution of the assignment for draw (d, c, i) is*

$$p(z_{dci} = k \mid \textit{rest}) \propto (\alpha_k + n_{dk}^{-d,c,i}) \cdot \frac{\beta_{kcw_{dci}} + n_{kcw_{dci}}^{-d,c,i}}{\sum_v \left(\beta_{kcv} + n_{kcv}^{-d,c,i}\right)}. \tag{A.2}$$

The first factor is the Pólya-urn predictive distribution of a Dirichlet-multinomial model. The second is the usual Dirichlet-multinomial predictive for the observed level. Under transfer, replace β_{kcv} in Equation (A.2) with $\tilde{\beta}_{kcv} = \beta_{kcv} + \lambda s_{kcv}$; the assignment formula is otherwise unchanged.

A.2 HDP GIBBS SAMPLING WITH HYPERPARAMETER UPDATES AND NO TRANSFER

The finite HDP approximation replaces the fixed document-topic prior by a shared global topic distribution and therefore augments the collapsed state with table counts and the concentrations τ and γ.

Corollary A.1 (HDP assignment update)**.** *In the finite HDP approximation, where document d has topic proportions distributed as $\mathrm{Dir}(\tau \boldsymbol{m})$, the conditional distribution of the assignment for draw (d, c, i) is*

$$p(z_{dci} = k \mid \textit{rest}) \propto (\tau m_k + n_{dk}^{-d,c,i}) \cdot \frac{\beta_{kcw_{dci}} + n_{kcw_{dci}}^{-d,c,i}}{\sum_v \left(\beta_{kcv} + n_{kcv}^{-d,c,i}\right)}. \tag{A.3}$$

The HDP changes only the document-level predictive factor. It lets the model share a global topic inventory while allowing each document to use a sparse subset of it; the emission calculation is unchanged. This is the standard Pólya-urn representation of a DP (Antoniak, 1974; Teh et al., 2006).

Proposition A.2 (Chinese restaurant franchise table update)**.** *In an HDP, suppose n_{dk} observations in document d are assigned to topic k. Conditional on τm_k, the number of tables t_{dk} can be sampled exactly by processing customers $i = 0, \dots, n_{dk} - 1$ and starting a new table with probability*

$$p(\textit{new table} \mid i, d, k) = \frac{\tau m_k}{\tau m_k + i}. \tag{A.4}$$

Thus, with independent $U_i \sim \mathrm{Uniform}(0,1)$,

$$t_{dk} = \sum_{i=0}^{n_{dk}-1} \mathbf{1}\left[U_i < \frac{\tau m_k}{\tau m_k + i}\right]. \tag{A.5}$$

This sequential construction is the standard Antoniak distribution and is the Chinese restaurant franchise representation of the HDP (Antoniak, 1974; Teh et al., 2006).

Corollary A.2 (HDP global-weight update)**.** *After sampling the table counts, let $q_k = \sum_d t_{dk}$ be the number of tables serving topic k. In a finite truncation with K global topics and symmetric base distribution γ/K, the global topic weights have the conditional distribution*

$$\boldsymbol{m} \mid \boldsymbol{q}, \gamma \sim \mathrm{Dirichlet}\left(q_1 + \frac{\gamma}{K}, \ldots, q_K + \frac{\gamma}{K}\right). \tag{A.6}$$

The table counts, rather than the number of observations assigned to each topic, are the observations seen by the global DP. The sampler draws the full Dirichlet vector, retaining posterior uncertainty in the global topic frequencies (Teh et al., 2006).

Proposition A.3 (Escobar & West (1995) concentration update without transfer)**.** *Consider G nonempty Chinese restaurant process groups with fixed sizes $n_g > 0$ and a shared concentration a. Let $\mathcal{P}_g$ be the partition of the n_g observations in group g into q_g occupied tables. The unaugmented model is*

$$a \sim \mathrm{Gamma}(a_0, b_0),$$
$$p(\mathcal{P}_g \mid n_g, a) = \frac{\Gamma(a)}{\Gamma(a+n_g)} a^{q_g} \prod_{T \in \mathcal{P}_g} \Gamma(|T|), \qquad g = 1, \ldots, G,$$

where $a_0, b_0 > 0$, the Gamma prior uses the shape–rate parameterization, and $|T|$ is the number of observations at table T. To sample a given the partitions $\{\mathcal{P}_g\}$, introduce for each group

$$\eta_g \mid a, n_g \sim \mathrm{Beta}(a+1, n_g), \tag{A.7}$$

$$\zeta_g \mid a, n_g \sim \mathrm{Bernoulli}\left(\frac{n_g}{n_g + a}\right). \tag{A.8}$$

Then the auxiliary-variable conditional is

$$a \mid \{\mathcal{P}_g, \eta_g, \zeta_g\}_{g=1}^{G} \sim \mathrm{Gamma}\left(a_0 + \sum_g q_g - \sum_g \zeta_g,\ b_0 - \sum_g \log \eta_g\right). \tag{A.9}$$

Corollary A.3 (Document concentration update)**.** *Let $\tau \sim \mathrm{Gamma}(a_\tau, b_\tau)$ in shape–rate parameterization. For each nonempty document d, let $n_d = \sum_k n_{dk} > 0$ be its number of topic assignments and $q_d = \sum_k t_{dk}$ its number of document-level tables. Draw independently across these documents*

$$\eta_d \mid \tau, n_d \sim \mathrm{Beta}(\tau + 1, n_d), \tag{A.10}$$

$$\zeta_d \mid \tau, n_d \sim \mathrm{Bernoulli}\left(\frac{n_d}{n_d + \tau}\right). \tag{A.11}$$

Then the auxiliary-variable update is

$$\tau \mid \{t_{dk}\}_{d,k}, \{\eta_d, \zeta_d\}_d \sim \mathrm{Gamma}\left(a_\tau + \sum_d q_d - \sum_d \zeta_d,\ b_\tau - \sum_d \log \eta_d\right), \tag{A.12}$$

where all sums run over nonempty documents. If every document is empty, draw τ from its prior.

Here each document is a group in Proposition A.3: its observations are customers, its t_{dk} are the tables serving topic k, and its total q_d is the group table count.

Corollary A.4 (Global concentration update without transfer)**.** *In the infinite-HDP Chinese restaurant franchise, let $\gamma \sim \mathrm{Gamma}(a_\gamma, b_\gamma)$ in shape–rate parameterization. The global restaurant has $Q = \sum_k q_k$ document-level tables as customers, where $q_k = \sum_d t_{dk}$, and $K_{\mathrm{act}} = |\{k : q_k > 0\}|$ occupied tables corresponding to active global topics. If $Q > 0$, draw*

$$\eta_0 \mid \gamma, Q \sim \mathrm{Beta}(\gamma + 1, Q), \tag{A.13}$$

$$\zeta_0 \mid \gamma, Q \sim \mathrm{Bernoulli}\left(\frac{Q}{Q + \gamma}\right). \tag{A.14}$$

Then the auxiliary-variable update is

$$\gamma \mid \boldsymbol{q}, \eta_0, \zeta_0 \sim \mathrm{Gamma}(a_\gamma + K_{\mathrm{act}} - \zeta_0,\ b_\gamma - \log \eta_0)\,. \tag{A.15}$$

If $Q = 0$, draw γ from its prior.

Here the global restaurant is the single group in Proposition A.3: its customers are document-level tables and its occupied tables are active topics. The update is exact for the infinite HDP but is the usual truncation approximation for the finite symmetric Dirichlet model used here. The exact finite-K alternative, including the no-transfer case $\lambda = 0$, is Proposition B.6. The underlying construction is due to Escobar & West (1995), with its HDP use described by Teh et al. (2006).

Procedure A.1 (One LDA-HDP collapsed Gibbs sweep)**.** *One complete sweep is:*

1. *Update every topic assignment z_{dci} using Corollary A.1.*
2. *Sample the document-topic table counts t_{dk} with Proposition A.2.*
3. *Update τ using the document restaurants in Corollary A.3.*
4. *Update γ using the global restaurant in Corollary A.4 (the finite-K truncation approximation).*
5. *Sample the global weights with the new γ using Corollary A.2.*

B Auxiliary-variable Gibbs updates under transfer

The exact Gibbs updates for γ and λ use the same collapsed Dirichlet–multinomial likelihood and the same beta, CRT, and binomial augmentations. We state these shared results first, combine them into a single Gamma update for a concentration parameter entering linearly (Proposition B.5), and obtain both updates as special cases.

B.1 Shared augmentation results

Proposition B.1 (Collapsed Dirichlet–multinomial block)**.** *Let $\boldsymbol{x} = (x_1, \dots, x_P)$, $X = \sum_j x_j$, and $\boldsymbol{a} = (a_1, \dots, a_P)$ with $a_j > 0$ and $A = \sum_j a_j$. If $\boldsymbol{\theta} \sim \mathrm{Dirichlet}(\boldsymbol{a})$ and $\boldsymbol{x} \mid X, \boldsymbol{\theta} \sim \mathrm{Multinomial}(X, \boldsymbol{\theta})$, then*

$$\begin{aligned} p(\boldsymbol{x} \mid X, \boldsymbol{a}) &= \frac{X!}{\prod_j x_j!} \frac{\mathrm{B}(\boldsymbol{a} + \boldsymbol{x})}{\mathrm{B}(\boldsymbol{a})} = \frac{X!}{\prod_j x_j!} \frac{\Gamma(A)}{\Gamma(A + X)} \prod_{j=1}^{P} \frac{\Gamma(a_j + x_j)}{\Gamma(a_j)} \\ &=: \mathrm{DirMult}(\boldsymbol{x} \mid X, \boldsymbol{a}). \end{aligned} \tag{B.1}$$

Hence its parameter-dependent likelihood kernel is the gamma-function ratio in Equation (B.1).

Proof. Integrating $\boldsymbol{\theta}$ over the simplex gives $\mathrm{B}(\boldsymbol{a} + \boldsymbol{x})/\mathrm{B}(\boldsymbol{a})$; expanding both multivariate beta functions gives the second equality. □

Proposition B.2 (Dirichlet–multinomial beta augmentation; Wang et al. (2024))**.** *Let $\boldsymbol{x}$, $X > 0$, $\boldsymbol{a}$, and A be as in Proposition B.1. Augmenting with*

$$u \mid \boldsymbol{x}, \boldsymbol{a} \sim \mathrm{Beta}(X, A) \tag{B.2}$$

gives

$$\mathrm{DirMult}(\boldsymbol{x} \mid X, \boldsymbol{a})\, p(u \mid \boldsymbol{x}, \boldsymbol{a}) = \frac{X!}{\Gamma(X) \prod_j x_j!}\, u^{X-1} (1-u)^{A-1} \prod_{j=1}^{P} \frac{\Gamma(a_j + x_j)}{\Gamma(a_j)}, \tag{B.3}$$

that is, the beta normalizing constant cancels the ratio $\Gamma(A)/\Gamma(A + X)$.

This is Lemma 2 of Wang et al. (2024); the identity follows by multiplying Equation (B.1) with the beta density.

Proposition B.3 (Chinese restaurant table augmentation (CRT); Antoniak (1974); Zhou & Carin (2015))**.** *Let $n \in \mathbb{N}_0$ be an observed count and $a > 0$ a concentration. Write the factor $\Gamma(a + n)/\Gamma(a)$ as the rising factorial $a^{\overline{n}} := \prod_{i=0}^{n-1} (a + i)$, with $a^{\overline{0}} := 1$, and let $\left[{n \atop \ell}\right]$ be the unsigned Stirling number of the first kind, with $\left[{0 \atop 0}\right] = 1$. The expansion*

$$a^{\overline{n}} = \sum_{\ell=0}^{n} \left[{n \atop \ell}\right] a^{\ell} \tag{B.4}$$

normalizes the auxiliary law

$$\ell \mid n, a \sim \mathrm{CRT}(n, a), \qquad p(\ell \mid n, a) = \frac{\left[{n \atop \ell}\right] a^{\ell}}{a^{\overline{n}}}, \tag{B.5}$$

so that $\frac{\Gamma(a+n)}{\Gamma(a)}\, p(\ell \mid n, a) = \left[{n \atop \ell}\right] a^{\ell}$. *Equivalently,* $\ell = \sum_{i=0}^{n-1} \xi_i$ *with independent* $\xi_i \sim \mathrm{Bernoulli}(a/(a+i))$*, the same sequential construction as in Proposition A.2.*

The table-count law is due to Antoniak (1974); the CRT terminology and the count augmentation used here follow Zhou & Carin (2015).

Proposition B.4 (Binomial attribution of an additive mass)**.** *Let* $a = \mu + \nu$ *with* $\mu, \nu \geq 0$ *and* $a > 0$*. For a nonnegative integer* ℓ*, augment the factor* a^{ℓ} *with*

$$g \mid \ell, \mu, \nu \sim \mathrm{Binomial}\left(\ell, \frac{\nu}{\mu + \nu}\right), \qquad \textit{so that} \qquad a^{\ell}\, p(g \mid \ell, \mu, \nu) = \binom{\ell}{g} \mu^{\ell - g} \nu^{g}, \tag{B.6}$$

with $0^0 = 1$.

Summing the right-hand side over $g = 0, \ldots, \ell$ recovers $(\mu + \nu)^{\ell}$ by the binomial theorem.

Proposition B.5 (Gamma update for a linearly entering concentration)**.** *Consider a finite set* $\mathcal{E}$ *of blocks and, for each block* $e \in \mathcal{E}$*, a count vector* $\boldsymbol{x}_e = (x_{e1}, \ldots, x_{eP_e})$ *with fixed total* $X_e = \sum_j x_{ej} > 0$ *and concentrations that depend linearly on a parameter* $\theta > 0$*,*

$$a_{ej}(\theta) = \mu_{ej} + \theta \nu_{ej} > 0, \qquad \mu_{ej}, \nu_{ej} \geq 0, \qquad \bar{\mu}_e = \sum_j \mu_{ej}, \qquad \bar{\nu}_e = \sum_j \nu_{ej}, \tag{B.7}$$

where μ_{ej} *and* ν_{ej} *do not depend on* θ*. Consider the model*

$$\theta \sim \mathrm{Gamma}(a_\theta, b_\theta), \tag{B.8}$$

$$\boldsymbol{\psi}_e \mid \theta \overset{\text{ind}}{\sim} \mathrm{Dirichlet}(a_{e1}(\theta), \ldots, a_{eP_e}(\theta)), \qquad e \in \mathcal{E}, \tag{B.9}$$

$$\boldsymbol{x}_e \mid X_e, \boldsymbol{\psi}_e \sim \mathrm{Multinomial}(X_e, \boldsymbol{\psi}_e), \qquad e \in \mathcal{E}, \tag{B.10}$$

where the Gamma prior uses the shape–rate parameterization. To sample θ *given the counts with the probability vectors* $\boldsymbol{\psi}_e$ *integrated out, draw for each block*

$$u_e \mid \theta, \boldsymbol{x}_e \sim \mathrm{Beta}(X_e, \bar{\mu}_e + \theta \bar{\nu}_e), \tag{B.11}$$

$$\ell_{ej} \mid \theta, x_{ej} \sim \mathrm{CRT}(x_{ej}, \mu_{ej} + \theta \nu_{ej}), \tag{B.12}$$

$$h_{ej} \mid \ell_{ej}, \theta \sim \mathrm{Binomial}\left(\ell_{ej}, \frac{\theta \nu_{ej}}{\mu_{ej} + \theta \nu_{ej}}\right). \tag{B.13}$$

Then, conditional on these auxiliary variables,

$$\theta \mid \{u_e, \boldsymbol{\ell}_e, \boldsymbol{h}_e, \boldsymbol{x}_e\}_{e \in \mathcal{E}} \sim \mathrm{Gamma}\left(a_\theta + \sum_{e \in \mathcal{E}} \sum_j h_{ej},\ b_\theta - \sum_{e \in \mathcal{E}} \bar{\nu}_e \log(1 - u_e)\right). \tag{B.14}$$

Proof. Because the blocks are independent given θ and each auxiliary variable depends on a single block, it suffices to derive the contribution of one block and then multiply the contributions of all blocks. Fix a block $e \in \mathcal{E}$ and drop the index e to lighten notation.

First, we write down the collapsed likelihood of the block. Applying Proposition B.1 with concentrations $a_j(\theta) = \mu_j + \theta \nu_j$, whose sum is the total concentration $\bar{\mu} + \theta \bar{\nu}$, and omitting the multinomial coefficient, which does not depend on θ, gives

$$L(\theta) = \frac{\Gamma(\bar{\mu} + \theta \bar{\nu})}{\Gamma(\bar{\mu} + \theta \bar{\nu} + X)} \prod_{j=1}^{P} \frac{\Gamma(\mu_j + \theta \nu_j + x_j)}{\Gamma(\mu_j + \theta \nu_j)}. \tag{B.15}$$

The first ratio, which involves the total concentration, prevents this likelihood from having a Gamma kernel in θ.

Second, the beta augmentation removes that ratio. By Proposition B.2, multiplying Equation (B.15) by the density of the beta auxiliary u in Equation (B.11) cancels the two gamma functions of the total concentration:

$$L(\theta)\, p(u \mid \theta, \boldsymbol{x}) = \frac{u^{X-1}(1-u)^{\bar{\mu}-1}}{\Gamma(X)}\, \exp\{\theta \bar{\nu} \log(1-u)\} \prod_{j=1}^{P} \frac{\Gamma(\mu_j + \theta \nu_j + x_j)}{\Gamma(\mu_j + \theta \nu_j)}, \tag{B.16}$$

where we split $(1-u)^{\bar{\mu}+\theta\bar{\nu}-1} = (1-u)^{\bar{\mu}-1}(1-u)^{\theta\bar{\nu}}$ to isolate an exponential in θ. One gamma ratio remains for each entry j.

Third, the CRT augmentation turns each remaining gamma ratio into a power of its concentration. Applying Proposition B.3 with count $n = x_j$ and concentration $a = \mu_j + \theta\nu_j$, multiplying the j-th ratio in Equation (B.16) by the probability of the auxiliary table count ℓ_j in Equation (B.12) gives

$$\frac{\Gamma(\mu_j + \theta\nu_j + x_j)}{\Gamma(\mu_j + \theta\nu_j)}\, p(\ell_j \mid x_j, \mu_j + \theta\nu_j) = \begin{bmatrix} x_j \\ \ell_j \end{bmatrix} (\mu_j + \theta\nu_j)^{\ell_j}. \tag{B.17}$$

The likelihood now contains powers of sums rather than powers of θ.

Fourth, the binomial attribution separates the fixed mass μ_j from the parameter-dependent mass $\theta\nu_j$. Applying Proposition B.4 with $\mu = \mu_j$ and $\nu = \theta\nu_j$, multiplying Equation (B.17) by the probability of the attributed count h_j in Equation (B.13) gives

$$(\mu_j + \theta\nu_j)^{\ell_j}\, p(h_j \mid \ell_j, \theta) = \binom{\ell_j}{h_j} \mu_j^{\ell_j - h_j} \nu_j^{h_j}\, \theta^{h_j}. \tag{B.18}$$

Combining the three augmentations, the augmented likelihood of the block is

$$L(\theta)\, p(u \mid \theta, \boldsymbol{x}) \prod_{j=1}^{P} p(\ell_j \mid x_j, \mu_j + \theta\nu_j)\, p(h_j \mid \ell_j, \theta) = \kappa(u, \boldsymbol{\ell}, \boldsymbol{h})\, \theta^{\sum_j h_j} \exp\{\theta\bar{\nu}\log(1-u)\}, \tag{B.19}$$

where the factor

$$\kappa(u, \boldsymbol{\ell}, \boldsymbol{h}) = \frac{u^{X-1}(1-u)^{\bar{\mu}-1}}{\Gamma(X)} \prod_{j=1}^{P} \begin{bmatrix} x_j \\ \ell_j \end{bmatrix} \binom{\ell_j}{h_j} \mu_j^{\ell_j - h_j} \nu_j^{h_j} \tag{B.20}$$

does not depend on θ, with $0^0 = 1$.

Finally, we restore the block index e and multiply the augmented likelihoods Equation (B.19) of all blocks with the shape–rate prior $p(\theta) \propto \theta^{a_\theta - 1} e^{-b_\theta \theta}$:

$$p(\theta \mid \{u_e, \boldsymbol{\ell}_e, \boldsymbol{h}_e, \boldsymbol{x}_e\}_{e \in \mathcal{E}}) \propto \theta^{a_\theta + \sum_e \sum_j h_{ej} - 1} \exp\left\{ -\left[b_\theta - \sum_e \bar{\nu}_e \log(1-u_e) \right] \theta \right\}. \tag{B.21}$$

This is the Gamma kernel of Equation (B.14); its rate is positive because $0 < u_e < 1$ implies $-\log(1-u_e) > 0$. Since each auxiliary variable is drawn from its conditional distribution given θ and the counts, the joint distribution of θ and the auxiliaries has the original posterior of θ as its marginal, so alternating the auxiliary draws with the Gamma draw is a valid Gibbs step. □

Blocks with $\bar{\nu}_e = 0$ do not depend on θ and can be dropped, and entries with $x_{ej} = 0$ give $\ell_{ej} = h_{ej} = 0$.

B.2 Global concentration update

Under transfer, the symmetric global prior is augmented by the frozen source table summaries. Let $q_k = \sum_d t_{dk}^{(t)}$, $Q = \sum_k q_k$, and $R = \sum_k r_k$. The finite-truncation update in Corollary A.2 becomes

$$\boldsymbol{m}^{(t)} \mid \boldsymbol{q}, \lambda, \gamma \sim \mathrm{Dirichlet}\Big(q_1 + \lambda r_1 + \frac{\gamma}{K}, \ldots, q_K + \lambda r_K + \frac{\gamma}{K}\Big). \tag{B.22}$$

The standard Escobar–West update for the global concentration is invalid when any transferred table offset λr_k is nonzero, even when λ is fixed. The document concentration τ is unaffected and retains Equation (A.9). We derive the exact finite-K update directly from the collapsed likelihood of the global table totals.

Proposition B.6 (Exact Gibbs update for γ under transfer)**.** *In the Chinese restaurant franchise, each of the $Q = \sum_k q_k$ document-level tables of the target model selects a global topic from* $\mathrm{Categorical}(\boldsymbol{m}^{(t)})$*, and $q_k = \sum_d t_{dk}^{(t)}$ counts the tables selecting topic k. With K topics, fixed transferred table summaries $r_k \geq 0$ with total $R = \sum_k r_k$, and fixed transfer strength $\lambda \geq 0$, the global table model is*

$$\gamma \sim \mathrm{Gamma}(a_\gamma, b_\gamma), \tag{B.23}$$

$$\boldsymbol{m}^{(t)} \mid \gamma, \lambda \sim \mathrm{Dirichlet}\Big(\lambda r_1 + \frac{\gamma}{K}, \ldots, \lambda r_K + \frac{\gamma}{K}\Big), \tag{B.24}$$

$$\boldsymbol{q} \mid Q, \boldsymbol{m}^{(t)} \sim \mathrm{Multinomial}(Q, \boldsymbol{m}^{(t)}), \tag{B.25}$$

where the Gamma prior uses the shape–rate parameterization. To sample γ given $\boldsymbol{q}$ with $\boldsymbol{m}^{(t)}$ integrated out, suppose $Q > 0$ and draw

$$u \mid \gamma, \boldsymbol{q}, \lambda \sim \mathrm{Beta}(Q, \lambda R + \gamma), \tag{B.26}$$

$$\ell_k = \sum_{i=0}^{q_k - 1} \xi_{ki}, \qquad \xi_{ki} \overset{\text{ind}}{\sim} \mathrm{Bernoulli}\left(\frac{\lambda r_k + \gamma/K}{\lambda r_k + \gamma/K + i}\right), \tag{B.27}$$

$$g_k \mid \ell_k, \gamma, \lambda \sim \mathrm{Binomial}\left(\ell_k, \frac{\gamma/K}{\lambda r_k + \gamma/K}\right), \tag{B.28}$$

for $k = 1, \ldots, K$, with $\ell_k = g_k = 0$ when $q_k = 0$, and then sample

$$\gamma \mid u, \boldsymbol{g}, \boldsymbol{q}, \lambda \sim \mathrm{Gamma}\left(a_\gamma + \sum_{k=1}^{K} g_k,\ b_\gamma - \log(1-u)\right). \tag{B.29}$$

If $Q = 0$, draw γ from its prior.

Proof. First, we identify the likelihood of γ. In the Chinese restaurant franchise, each of the Q document-level tables selects its global topic independently from $\mathrm{Categorical}(\boldsymbol{m}^{(t)})$. Given $\boldsymbol{m}^{(t)}$, the probability of these selections is $\prod_{k=1}^{K} (m_k^{(t)})^{q_k}$, so the rest of the sampler state interacts with $\boldsymbol{m}^{(t)}$, and hence with γ, only through the table totals $\boldsymbol{q}$. Integrating out $\boldsymbol{m}^{(t)} \sim \mathrm{Dirichlet}(\lambda r_1 + \gamma/K, \ldots, \lambda r_K + \gamma/K)$ and applying Proposition B.1 with counts $x_k = q_k$ and concentrations $a_k = \lambda r_k + \gamma/K$ gives a Dirichlet–multinomial likelihood for $\boldsymbol{q}$ with total concentration $\lambda R + \gamma$.

Second, we match this likelihood to Proposition B.5. Each concentration is the sum of a fixed part and a part proportional to γ,

$$a_k(\gamma) = \underbrace{\lambda r_k}_{\mu_k} + \gamma \cdot \underbrace{\tfrac{1}{K}}_{\nu_k}, \qquad \bar{\mu} = \sum_k \mu_k = \lambda R, \qquad \bar{\nu} = \sum_k \nu_k = 1, \tag{B.30}$$

so the table totals form a single block of Proposition B.5 with parameter $\theta = \gamma$, counts $\boldsymbol{x} = \boldsymbol{q}$, and total $X = Q > 0$.

Third, we read off the auxiliary variables. With these values, the beta auxiliary of Equation (B.11) is $\mathrm{Beta}(Q, \lambda R + \gamma)$, which is Equation (B.26). The CRT auxiliary of Equation (B.12) is $\ell_k \sim \mathrm{CRT}(q_k, \lambda r_k + \gamma/K)$, which Proposition B.3 samples by the Bernoulli construction in Equation (B.27). The attributed count of Equation (B.13) has success probability $(\gamma/K)/(\lambda r_k + \gamma/K)$, so it is g_k in Equation (B.28).

Finally, substituting $\bar{\nu} = 1$ and the attributed counts g_k into Equation (B.14) gives the conditional $\mathrm{Gamma}(a_\gamma + \sum_k g_k,\ b_\gamma - \log(1-u))$ of Equation (B.29). If $Q = 0$, every $q_k = 0$, the collapsed likelihood equals one, and the conditional is the prior. □

To summarize, the procedure to fit LDA-HDP models under transfer is as follows.

Procedure B.1 (One transfer LDA-HDP collapsed Gibbs sweep)**.** *One complete sweep is:*

1. *Update every topic assignment using Equation* (A.3) *and prior $\beta_{kcv} + \lambda s_{kcv}$.*
2. *Sample the table counts with Equation* (A.4) *and compute $q_k = \sum_d t_{dk}^{(t)}$ and $Q = \sum_k q_k$.*
3. *If $Q > 0$, update γ with equation B.26–equation B.29; otherwise draw it from its prior.*
4. *Update τ with Equation* (A.9)*.*
5. *Sample the global weights with the updated γ and current λ using Equation* (B.22)*.*

B.3 Transfer-strength update

We now derive an exact auxiliary-variable Gibbs update for the transfer strength λ applicable when estimating target topics. Unlike γ, λ occurs in both the global table likelihood and every transferred emission block, so each nonconstant block requires its own auxiliaries. The following proposition augments all of these blocks simultaneously.

Proposition B.7 (Exact Gibbs update for λ under transfer)**.** *Fix the global concentration $\gamma > 0$ and the source summaries $r_k, s_{kcv} \geq 0$, let q_k, Q, and R be as in Proposition B.6, and, for each topic–channel block (k, c), let*

$$B_{kc} = \sum_v \beta_{kcv}, \qquad S_{kc} = \sum_v s_{kcv}, \qquad N_{kc}^{(t)} = \sum_v n_{kcv}^{(t)}, \tag{B.31}$$

so that the total prior concentration of the block is $A_{kc}(\lambda) = B_{kc} + \lambda S_{kc}$ as in Equation (9). The transfer strength enters both the global table model of Proposition B.6 and every emission block:

$$\lambda \sim \mathrm{Gamma}(a_\lambda, b_\lambda), \tag{B.32}$$

$$\boldsymbol{m}^{(t)} \mid \lambda, \gamma \sim \mathrm{Dirichlet}\Big(\lambda r_1 + \frac{\gamma}{K}, \ldots, \lambda r_K + \frac{\gamma}{K}\Big), \tag{B.33}$$

$$\boldsymbol{q} \mid Q, \boldsymbol{m}^{(t)} \sim \mathrm{Multinomial}(Q, \boldsymbol{m}^{(t)}), \tag{B.34}$$

$$\boldsymbol{\phi}_{kc}^{(t)} \mid \lambda \overset{\text{ind}}{\sim} \mathrm{Dirichlet}(\boldsymbol{\beta}_{kc} + \lambda \boldsymbol{s}_{kc}), \tag{B.35}$$

$$\boldsymbol{n}_{kc}^{(t)} \mid N_{kc}^{(t)}, \boldsymbol{\phi}_{kc}^{(t)} \sim \mathrm{Multinomial}(N_{kc}^{(t)}, \boldsymbol{\phi}_{kc}^{(t)}), \tag{B.36}$$

independently across topic–channel blocks, where the Gamma prior uses the shape–rate parameterization. To sample λ given the table totals and the emission counts with $\boldsymbol{m}^{(t)}$ and all $\boldsymbol{\phi}_{kc}^{(t)}$ integrated out, let $\mathcal{I} = \{(k,c) : N_{kc}^{(t)} > 0,\ S_{kc} > 0\}$. For the global table block, if $Q > 0$ and $R > 0$, draw

$$u_0 \mid \lambda, \gamma, \boldsymbol{q} \sim \mathrm{Beta}(Q, \lambda R + \gamma), \tag{B.37}$$

$$\ell_k^{(0)} \mid \lambda, \gamma, q_k \sim \mathrm{CRT}\Big(q_k, \lambda r_k + \frac{\gamma}{K}\Big), \tag{B.38}$$

$$h_k^{(0)} \mid \ell_k^{(0)}, \lambda, \gamma \sim \mathrm{Binomial}\left(\ell_k^{(0)}, \frac{\lambda r_k}{\lambda r_k + \gamma/K}\right), \tag{B.39}$$

and set

$$H_0 = \sum_k h_k^{(0)}, \qquad C_0 = -R \log(1 - u_0); \tag{B.40}$$

otherwise set $H_0 = C_0 = 0$. For every $(k,c) \in \mathcal{I}$, draw

$$u_{kc} \mid \lambda, \boldsymbol{n}_{kc}^{(t)} \sim \mathrm{Beta}\Big(N_{kc}^{(t)}, A_{kc}(\lambda)\Big), \tag{B.41}$$

$$\ell_{kcv} \mid \lambda, n_{kcv}^{(t)} \sim \mathrm{CRT}\Big(n_{kcv}^{(t)}, \beta_{kcv} + \lambda s_{kcv}\Big), \tag{B.42}$$

$$h_{kcv} \mid \ell_{kcv}, \lambda \sim \mathrm{Binomial}\left(\ell_{kcv}, \frac{\lambda s_{kcv}}{\beta_{kcv} + \lambda s_{kcv}}\right), \tag{B.43}$$

and set

$$H_{\mathrm{em}} = \sum_{(k,c)\in\mathcal{I}} \sum_v h_{kcv}, \qquad C_{\mathrm{em}} = -\sum_{(k,c)\in\mathcal{I}} S_{kc} \log(1 - u_{kc}). \tag{B.44}$$

Then

$$\lambda \mid \textit{auxiliaries}, \boldsymbol{q}, \{\boldsymbol{n}_{kc}^{(t)}\}_{kc}, \gamma \sim \mathrm{Gamma}(a_\lambda + H_0 + H_{\mathrm{em}},\ b_\lambda + C_0 + C_{\mathrm{em}}). \tag{B.45}$$

Proof. First, we identify the likelihood of λ. With the global weights $\boldsymbol{m}^{(t)}$ and the emission vectors $\boldsymbol{\phi}_{kc}^{(t)}$ integrated out, λ appears in two kinds of factors. The table totals $\boldsymbol{q}$ have, as shown in the proof of Proposition B.6, a Dirichlet–multinomial likelihood with concentrations $\lambda r_k + \gamma/K$. For each topic–channel block (k,c), the emission counts $\boldsymbol{n}_{kc}^{(t)}$ given their total $N_{kc}^{(t)}$ have, by Proposition B.1, a Dirichlet–multinomial likelihood with concentrations $\beta_{kcv} + \lambda s_{kcv}$ and total concentration $A_{kc}(\lambda) = B_{kc} + \lambda S_{kc}$. The global weights and the emission vectors are a priori independent, so the collapsed likelihood is the product of the table block and all emission blocks.

Second, we discard the blocks that do not depend on λ. The table block equals one when $Q = 0$ and does not depend on λ when $R = 0$. Likewise, an emission block equals one when $N_{kc}^{(t)} = 0$ and does not depend on λ when $S_{kc} = 0$. The remaining blocks are the table block when $Q > 0$ and $R > 0$, and the emission blocks $(k,c) \in \mathcal{I}$.

Third, we match each remaining block to Proposition B.5 with parameter $\theta = \lambda$. For the table block, the concentrations and their sums are

$$a_k(\lambda) = \underbrace{\gamma/K}_{\mu_k} + \lambda \underbrace{r_k}_{\nu_k}, \qquad \bar{\mu} = \gamma, \qquad \bar{\nu} = R, \tag{B.46}$$

with counts $\boldsymbol{x} = \boldsymbol{q}$ and total $X = Q$. The beta, CRT, and attribution auxiliaries of Proposition B.5 are then equation B.37–equation B.39, and the block contributes $H_0 = \sum_k h_k^{(0)}$ to the shape and

$C_0 = -R\log(1-u_0)$ to the rate, as defined in Equation (B.40). For each emission block $(k,c) \in \mathcal{I}$, the concentrations and their sums are

$$a_v(\lambda) = \underbrace{\beta_{kcv}}_{\mu_v} + \lambda \underbrace{s_{kcv}}_{\nu_v}, \qquad \bar{\mu} = B_{kc}, \qquad \bar{\nu} = S_{kc}, \tag{B.47}$$

with counts $\boldsymbol{x} = \boldsymbol{n}_{kc}^{(t)}$ and total $X = N_{kc}^{(t)}$. The auxiliaries of Proposition B.5 are then equation B.41–equation B.43, and the block contributes $\sum_v h_{kcv}$ to the shape and $-S_{kc}\log(1-u_{kc})$ to the rate. Summing over $(k,c) \in \mathcal{I}$ gives H_{em} and C_{em} in Equation (B.44).

Finally, Equation (B.14) applied to all included blocks gives shape $a_\lambda + H_0 + H_{\mathrm{em}}$ and rate $b_\lambda + C_0 + C_{\mathrm{em}}$, which is Equation (B.45). When the table block is omitted, $H_0 = C_0 = 0$, consistent with the definition in the statement. □

The γ and λ updates use fresh auxiliary variables and condition sequentially on the current value of the other parameter. Both integrate out $\boldsymbol{m}^{(t)}$ even though the table counts were sampled conditional on it; this is a partially collapsed Gibbs sampler, which leaves the posterior invariant provided $\boldsymbol{m}^{(t)}$ is redrawn from its full conditional immediately after the collapsed updates (van Dyk & Park, 2008). The global weights are therefore resampled only after both concentration updates.

Procedure B.2 (Complete HDP Gibbs sweep under transfer)**.** *Up to optional monitoring and posterior accumulation, one complete sweep is:*

1. *Sample z_{dci} using its conditional in Equation* (A.3) *and prior $\beta_{kcv} + \lambda s_{kcv}$ for each draw (d,c,i).*
2. *For every (d,k), sample the table count $t_{dk}^{(t)}$ by the Bernoulli construction in Equation* (A.4)*; then set $q_k = \sum_d t_{dk}^{(t)}$ and $Q = \sum_k q_k$.*
3. *Update γ with the auxiliary variables and Gamma conditional in Proposition B.6.*
4. *Update the document concentration τ with the Escobar–West auxiliary-variable update in Equation* (A.9)*.*
5. *Update λ with the global-table and emission auxiliaries and Gamma conditional in Proposition B.7, using the current γ and the same table totals as Step 3.*
6. *With the updated γ and λ, sample $\boldsymbol{m}$ using Equation* (B.22)*.*

C Experimental details

C.1 Model fitting

Priors for γ, τ and λ were $\mathrm{Gamma}(1,1) = \mathrm{Exp}(1)$. Models fit without transfer (source-only, target-only, and joint fits) sampled γ with the Escobar–West update of Corollary A.4, while transfer models used the exact auxiliary-variable updates in Propositions B.6 and B.7, which require no tuning. For the Cleveland dataset, we used $K = 12$ topics with 100,000 warmup and 100,000 sampling steps with thinning of 200 to obtain 500 posterior samples, while for the synthetic data we used the true number of topics $K = 4$ with 20,000 steps for warmup and 20,000 for sampling with thinning of 50 for 400 posterior samples. The grid search oracle fit transfer models for each $\lambda \in \{0.01, 0.03, 0.1, 0.3, 0.5, 1.0, 2.0, 3.0, 10.0, 30.0\}$ and, for each dataset configuration, selected the fixed λ with lowest average distance to the reference topics across repetitions. To find $\hat{\lambda}$ in Equation (12), we computed the nominal covariances in Equation (9) for 500 choices of λ in a log-spaced grid between 10^{-3} and 10^{1} and chose the value with minimum median squared Frobenius norm; Dirichlet MLE used the same grid.

C.2 Alternative heuristics for the transfer strength

Proposition C.1 (Dirichlet MLE rule)**.** *Let $\tilde{\boldsymbol{\phi}}_{kc}^{(s,j)}$ be the source posterior draws of Equation* (10)*. The Dirichlet MLE transfer strength is the value for which the transfer prior assigns the highest density to these draws, aggregating over topic-channel blocks with the median as in Equation* (12)*:*

$$\hat{\lambda}_{\mathrm{MLE}} = \arg\max_{\lambda>0} \mathrm{Median}_{(k,c)} \ \frac{1}{J}\sum_{j=1}^{J} \log \mathrm{Dir}\Big(\tilde{\boldsymbol{\phi}}_{kc}^{(s,j)} \mid \boldsymbol{\beta}_{kc} + \lambda \boldsymbol{s}_{kc}\Big). \tag{C.1}$$

Proposition C.2 (Target-mass matching rule)**.** *Following the notation of Definition 2.1, where N_{dc} is the number of observed categorical draws for channel c in document d and missing observations are not counted, define the total token counts of the source and target datasets as*

$$T^{(s)} = \sum_{d=1}^{D^{(s)}} \sum_{c=1}^{C} N_{dc}^{(s)}, \qquad T^{(t)} = \sum_{d=1}^{D^{(t)}} \sum_{c=1}^{C} N_{dc}^{(t)}. \tag{C.2}$$

The target-mass matching transfer strength is

$$\lambda_{\mathrm{TM}} = \frac{T^{(t)}}{T^{(s)}}. \tag{C.3}$$

Since $\sum_{k,c} S_{kc} = T^{(s)}$, the scaled source counts then carry the same total mass as the target counts, $\lambda_{\mathrm{TM}} \sum_{k,c} S_{kc} = T^{(t)}$.

C.3 Synthetic Dataset Generation

Fixing the number of topics K and channels C, we first generate emission-base vectors for each topic k and channel c with size V_c as $\boldsymbol{\omega}_{kc} \sim \mathrm{Dir}(\alpha_0 \mathbf{1}_{V_c})$. The base concentration α_0 controls how peaked or diffuse each topic is. The emission bases are used to sample population-level source and target topics $\boldsymbol{\phi}_{kc}^{(s)} \sim \mathrm{Dir}(\kappa^{(s)} \boldsymbol{\omega}_{kc})$ and $\boldsymbol{\phi}_{kc}^{(t)} \sim \mathrm{Dir}(\kappa^{(t)} \boldsymbol{\omega}_{kc})$ where high values of $\kappa^{(s)}$ and $\kappa^{(t)}$ encourage topics to cluster around $\boldsymbol{\omega}_{kc}$ and low values encourage more spread. The population joint topics are computed as $(D^{(s)}\boldsymbol{\phi}^{(s)} + D^{(t)}\boldsymbol{\phi}^{(t)})/(D^{(s)} + D^{(t)})$. We quantify the domain shift as $\bar{\Delta}^2 = \frac{1}{K}\sum_{k=1}^{K}\sum_{c=1}^{C} \|\boldsymbol{\phi}_{kc}^{(s)} - \boldsymbol{\phi}_{kc}^{(t)}\|_2^2$ and group dataset pairs into its quartiles. We then sample topic vectors for source and target documents as $\boldsymbol{\pi}_d^{(s)} \sim \mathrm{Dir}(\alpha^{(s)} \mathbf{1}_K)$ and $\boldsymbol{\pi}_d^{(t)} \sim \mathrm{Dir}(\alpha^{(t)} \mathbf{1}_K)$, where $\alpha^{(s)}$ and $\alpha^{(t)}$ control how many topics are used in each document, and from there latent topics and observations as described previously. We used $K = 4$ latent topics, $C = 6$ channels with dimensionality $V_c = 5$, resulting in a total of 120 parameters to estimate; if we demand at least 15 samples per parameter, at least 300 documents are needed to properly estimate the topic signatures. Therefore, we used $D^{(s)}, D^{(t)} \in \{100, 1000\}$ source and target documents covering both well-estimated and poorly-estimated emissions. We sweep $\alpha_0, \alpha^{(s)}, \alpha^{(t)} \in \{0.1, 0.5, 1.0\}$ covering peaked and well-separated to flat and overlapping topics (α_0) and few to many topics used for each document ($\alpha^{(s)}, \alpha^{(t)}$), $\kappa^{(s)}, \kappa^{(t)} \in \{3, 30, 300\}$ ranging from large to small domain shift.

C.4 Cleveland dataset preprocessing

We used the comprehensive Cleveland heart-disease dataset retrieved from OpenML (data ID 43672), which combines the Cleveland, Hungarian, Switzerland, Long Beach VA, and Statlog heart disease cohorts. The input consisted of six categorical features: sex, chest-pain type, fasting blood sugar, resting electrocardiographic result, exercise-induced angina, and ST-segment slope, and five continuous features: age, resting blood pressure, serum cholesterol, maximum heart rate, and old-peak. The disease target was retained separately and was used only to define the source and target groups in the out-of-distribution experiments; it was not included as an input feature.

Categorical features were converted to integer category codes using their observed levels. Known sentinels for missing and invalid values in the original encoding were replaced with -1 which the sampler interprets as a missing value. Specifically, a value of zero for ST-segment slope, values below 75 for resting blood pressure or cholesterol, and negative values of oldpeak, were treated as invalid and replaced with -1.

For the continuous features, we did not apply z-score normalization, min-max scaling, clipping, or imputation, we instead discretized them into eight categories based on empirical quantiles. Quantile boundaries were computed from the complete dataset before splitting into source and target datasets. This is not a form of data leakage, but rather a necessary step to ensure that categories in both datasets are aligned correctly and refer to the same biomarker ranges. The six original categorical channels and the five discretized continuous channels were concatenated to form the model input.

C.5 ADDITIONAL RESULTS ON THE SYNTHETIC DATASETS

Estimand	Method	Domain shift Dataset sizes (src/tgt)	(0.0, 0.12]	(0.12, 0.38]	(0.38, 0.93]	(0.93, 2.58]
Joint Topics	Best Grid Search	100 / 100	0.21(1)	0.15(1)	0.22(1)	0.190(9)
		100 / 1000	0.15(1)	0.139(7)	0.116(9)	0.109(6)
		1000 / 100	2.0(3)	1.1(2)	3.5(4)	3.7(4)
		1000 / 1000	0.22(1)	0.17(1)	0.34(2)	0.29(2)
Target Topics	Adaptive Gibbs	100 / 100	0.69(2)	0.43(2)	0.133(4)	0.074(2)
		100 / 1000	0.30(1)	0.196(7)	0.083(2)	0.054(1)
		1000 / 100	0.73(2)	0.28(2)	0.0176(5)	0.0090(3)
		1000 / 1000	0.22(2)	0.048(2)	0.0132(3)	0.0070(2)
	Best Grid Search	100 / 100	0.152(9)	0.074(4)	0.045(2)	0.035(1)
		100 / 1000	0.122(7)	0.117(6)	0.043(3)	0.057(3)
		1000 / 100	1.6(3)	0.026(1)	0.0117(3)	0.001(0)
		1000 / 1000	0.061(4)	0.025(1)	0.0151(5)	0.0114(3)

Table C.1: Final lambda values for different methods, domain shifts, and dataset sizes for the synthetic datasets. Standard error on the mean as shorthand notation.

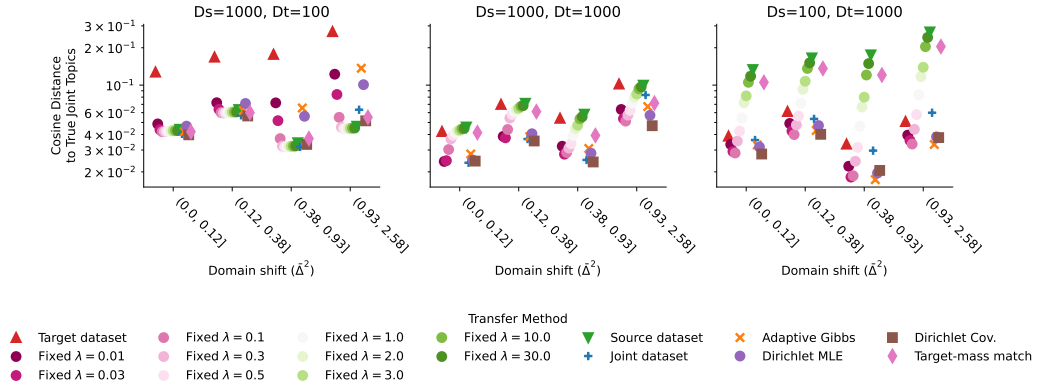


Figure C.1: Disaggregated data for Figure 1 (left) when estimating joint topics, excluding $D_s = D_t = 100$ for clarity. Cosine distance to true joint topics (y-axis) by domain shift $\bar{\Delta}^2$ (x-axis, Appendix C.3) for different transfer methods (hue colors).

Estimand	Domain shift ($\bar{\Delta}^2$) Method	(0.0, 0.12]	(0.12, 0.38]	(0.38, 0.93]	(0.93, 2.58]
Joint topics	Adaptive Gibbs	0.049(2)	0.064(2)	0.048(1)	0.089(2)
	Best Grid Search	0.040(1)	0.056(2)	0.035(1)	0.056(1)
	Joint dataset	0.045(2)	0.065(2)	0.040(2)	0.088(2)
	Dirichlet MLE	0.047(2)	0.069(2)	0.047(1)	0.082(2)
	Dirichlet Cov.	0.043(1)	0.059(2)	0.039(1)	0.063(1)
	Target-mass match	0.073(2)	0.096(2)	0.074(2)	0.118(2)
	Source dataset	0.088(2)	0.113(2)	0.100(2)	0.148(3)
	Target dataset	0.083(2)	0.113(2)	0.099(2)	0.150(3)
Target topics	Adaptive Gibbs	0.051(2)	0.069(2)	0.045(2)	0.073(2)
	Best Grid Search	0.041(1)	0.060(2)	0.042(1)	0.072(2)
	Joint dataset	0.047(2)	0.076(2)	0.074(2)	0.154(3)
	Dirichlet MLE	0.047(2)	0.068(2)	0.046(1)	0.083(2)
	Dirichlet Cov.	0.044(1)	0.066(2)	0.062(2)	0.115(2)
	Target-mass match	0.076(2)	0.109(2)	0.110(2)	0.205(2)
	Source dataset	0.093(2)	0.133(2)	0.159(2)	0.274(2)
	Target dataset	0.081(2)	0.106(2)	0.075(2)	0.098(2)

Table C.2: Average cosine distance to true joint (top block) and target (bottom block) topics for different methods and different domain shifts for the synthetic datasets. Standard error on the mean as shorthand notation (e.g., $1.23(4)$ denotes a mean of 1.23 with a standard error of 0.04).

Estimand	Dataset sizes source / target Method	100 / 100	100 / 1000	1000 / 100	1000 / 1000
Joint topics	Adaptive Gibbs	0.101(2)	0.032(1)	0.077(2)	0.041(1)
	Best Grid Search	0.090(2)	0.028(1)	0.041(1)	0.0285(8)
	Joint dataset	0.103(2)	0.045(2)	0.048(2)	0.043(1)
	Dirichlet MLE	0.105(2)	0.034(1)	0.069(2)	0.038(1)
	Dirichlet Cov.	0.093(2)	0.032(1)	0.045(1)	0.033(1)
	Target-mass match	0.118(2)	0.142(2)	0.049(1)	0.053(2)
	Source dataset	0.148(2)	0.187(3)	0.047(2)	0.069(2)
	Target dataset	0.147(2)	0.046(2)	0.186(2)	0.067(2)
Target Topics	Adaptive Gibbs	0.106(2)	0.032(1)	0.072(2)	0.028(1)
	Best Grid Search	0.093(2)	0.028(1)	0.069(2)	0.0251(9)
	Joint dataset	0.124(2)	0.046(2)	0.119(3)	0.065(2)
	Dirichlet MLE	0.102(2)	0.034(1)	0.078(2)	0.031(1)
	Dirichlet Cov.	0.105(2)	0.032(1)	0.102(2)	0.050(2)
	Target-mass match	0.153(2)	0.151(3)	0.101(2)	0.100(2)
	Source dataset	0.198(3)	0.198(3)	0.134(3)	0.134(3)
	Target dataset	0.134(2)	0.045(2)	0.135(2)	0.045(2)

Table C.3: Cosine distance to true joint (top block) and target (bottom block) topics for different methods and different dataset sizes for the synthetic datasets. Standard error on the mean as shorthand notation.

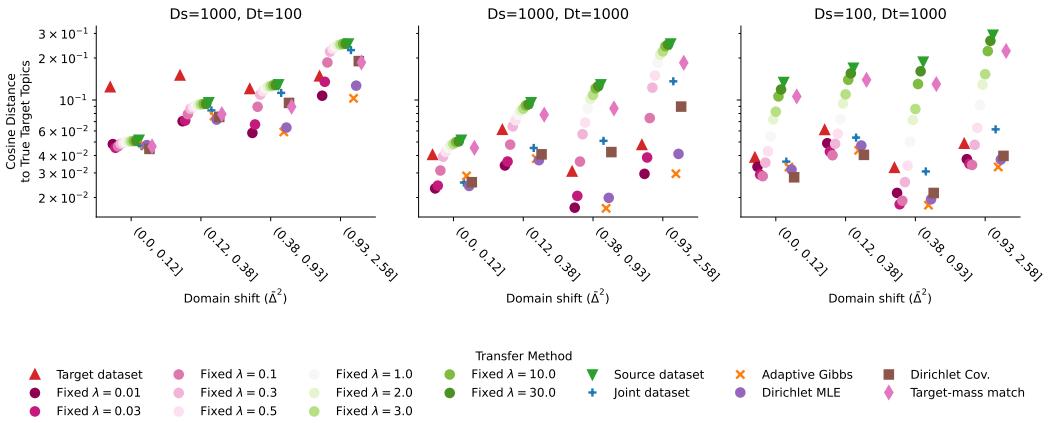


Figure C.2: Disaggregated data for Figure 1 (right) when estimating target topics, excluding $D_s = D_t = 100$ for clarity. Cosine distance to true target topics (y-axis) by domain shift $\bar{\Delta}^2$ (x-axis, Appendix C.3) for different transfer methods (hue colors).

C.6 Additional results on the Cleveland datasets

Hyperp.	$D^{(s)}/D^{(t)}$	Dir. Cov.	Adapt. Gibbs	Joint dataset	Source dataset	Target dataset	Grid
Fixed	0.10	0.095(5)	0.103(5)	0.055(4)	0.216(2)	0.081(4)	0.080(5)
	0.25	0.093(5)	0.086(4)	0.055(4)	0.206(2)	0.102(4)	0.082(4)
	0.50	0.107(5)	0.109(4)	0.055(4)	0.178(4)	0.131(5)	0.100(4)
	1.00	0.110(4)	0.104(5)	0.055(4)	0.152(5)	0.164(4)	0.106(4)
	2.00	0.099(3)	0.098(3)	0.055(4)	0.118(4)	0.199(3)	0.098(3)
	5.00	0.091(4)	0.093(4)	0.055(4)	0.102(4)	0.208(2)	0.089(4)
	10.00	0.093(4)	0.097(4)	0.055(4)	0.096(5)	0.225(2)	0.092(4)
Opt.	0.10	0.087(4)	0.087(4)	0.059(4)	0.243(2)	0.083(4)	0.077(5)
	0.25	0.102(5)	0.106(5)	0.059(4)	0.233(2)	0.092(4)	0.094(4)
	0.50	0.123(5)	0.124(5)	0.059(4)	0.205(4)	0.134(5)	0.120(5)
	1.00	0.128(4)	0.122(5)	0.059(4)	0.171(4)	0.168(5)	0.118(5)
	2.00	0.101(4)	0.097(3)	0.059(4)	0.114(4)	0.203(4)	0.097(3)
	5.00	0.094(3)	0.098(3)	0.059(4)	0.102(4)	0.224(3)	0.093(3)
	10.00	0.087(4)	0.091(4)	0.059(4)	0.089(5)	0.232(3)	0.085(4)

Table C.4: Data for Figure C.4. Cosine distances to joint dataset topics in the IID setting for the Cleveland dataset. Joint dataset baseline indicates variability between runs, and Grid indicates the best grid search results. Standard error on the mean as shorthand notation.

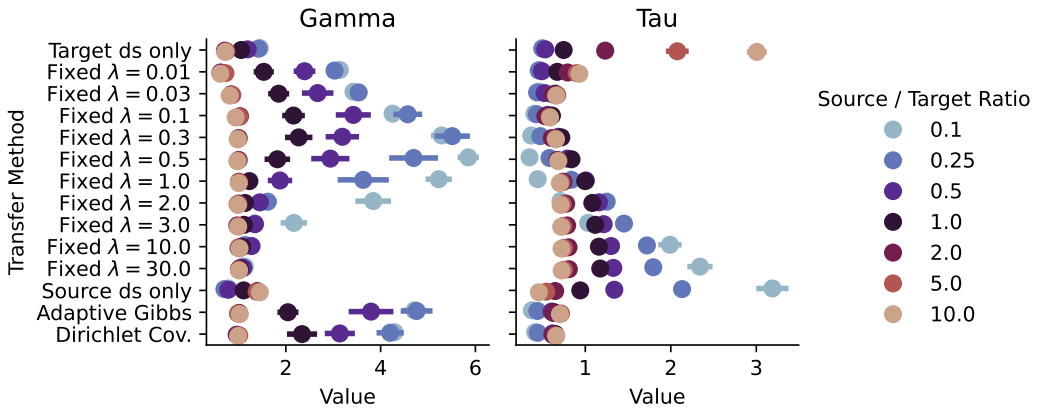


Figure C.3: Posterior averages of the HDP hyperparameters γ (left) and τ (right) when optimized on the target dataset for transfer learning in the IID setting.

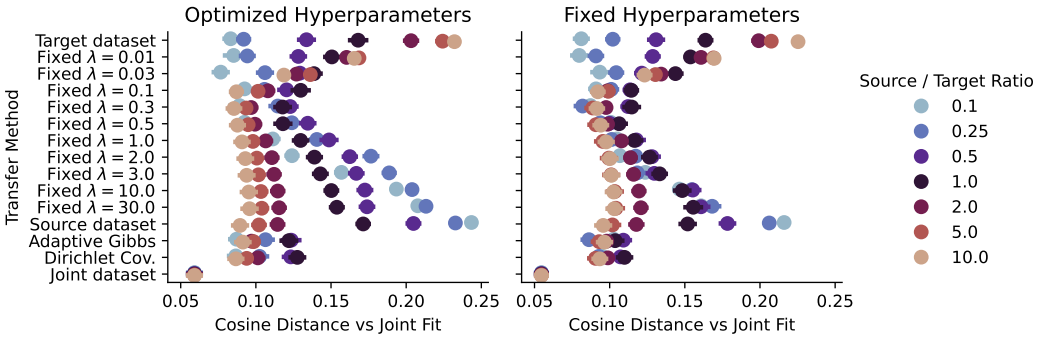


Figure C.4: Results on the IID setting for the Cleveland dataset for different values of λ and different relative sizes of source and target datasets with optimized hyperparameters (left) and hyperparameters fixed to the values found by the joint fit (right).

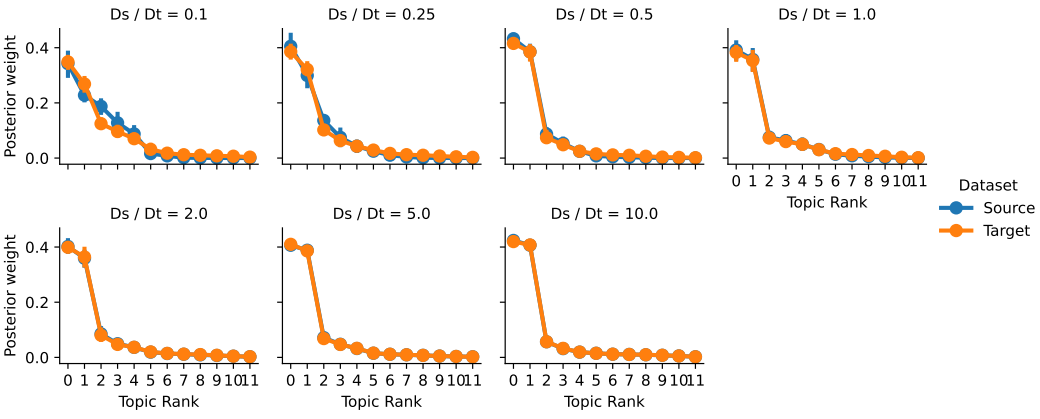


Figure C.5: Posterior topic weights $\boldsymbol{m}$ for the covariance matching heuristic in the Cleveland IID scenario when fitting on the source dataset (blue) and when fitting on the target dataset (orange) for different source to target dataset sizes (facets).

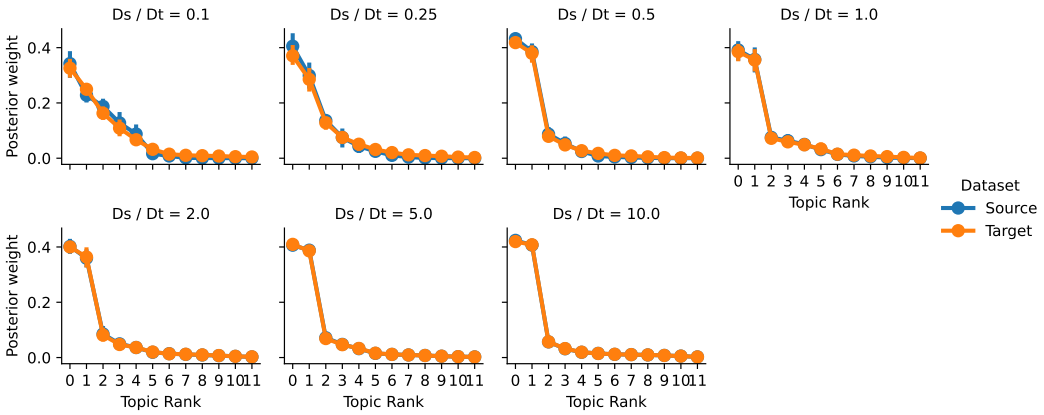


Figure C.6: Posterior topic weights $\boldsymbol{m}$ for fixed $\lambda = 0.3$ in the Cleveland IID scenario when fitting on the source dataset (blue) and when fitting on the target dataset (orange) for different source to target dataset sizes (facets).

Transfer Hyperp.	CVD to Healthy Fixed Hp.	CVD to Healthy Optimized Hp.	Healthy to CVD Fixed Hp.	Healthy to CVD Optimized Hp.
Dir. Cov.	0.118(4)	0.161(5)	0.128(5)	0.149(3)
Adapt. Gibbs	0.166(6)	0.171(5)	0.176(5)	0.182(3)
Grid	0.112(3)	0.156(5)	0.121(4)	0.139(4)
IID Equiv.	0.106(4)	0.118(5)	0.106(4)	0.118(5)
Joint dataset	0.071(8)	0.076(6)	0.060(5)	0.04(1)
Source dataset	0.171(6)	0.192(4)	0.177(3)	0.168(5)
Target dataset	0.177(3)	0.168(5)	0.171(6)	0.192(4)

Table C.5: Data for Figure C.7. Cosine distances to joint topics in the OOD setting for the Cleveland dataset. Joint dataset baseline indicates variability between runs. IID Equiv. indicates the grid search result with an 1:1 split of the source and target datasets achieved by the best performing λ and optimized hyperparameters. Standard error on the mean as shorthand notation. The performance of grid search and Dirichlet covariance was never statistically different across transfer direction and hyperparameter optimization ($p > 0.09$). Comparing with grid search with the equivalent IID split and fixed hyperparameters, the only statistically significant difference was with the Dirichlet covariance heuristic when transferring from Healthy to CVD ($p = 0.013$), while the other three comparisons were not significant.

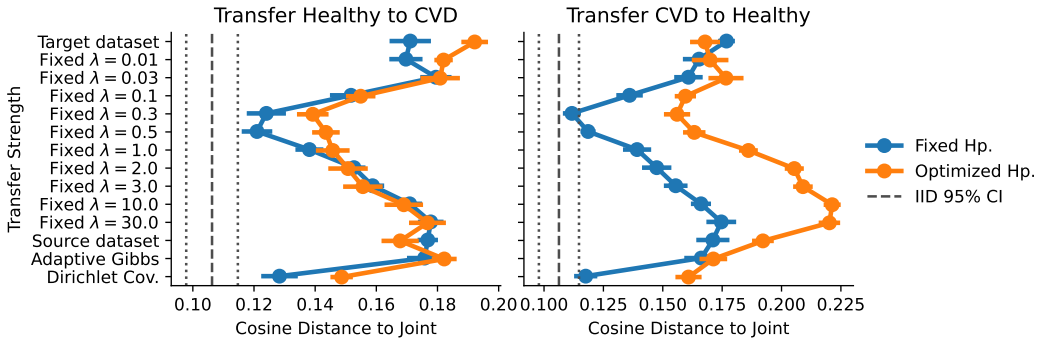


Figure C.7: Transfer learning results on the OOD scenario for the Cleveland dataset when transferring from Healthy to CVD (left) and vice-versa (right) for optimized (orange) hyperparameters, or hyperparameters fixed to the posterior averages found on the entire dataset (blue). Error bars denote the standard error of the mean. The black dashed line indicates the cosine distance of transfer models fit on IID splits of the dataset with roughly the same relative sizes as the OOD splits, along with 95% confidence interval.

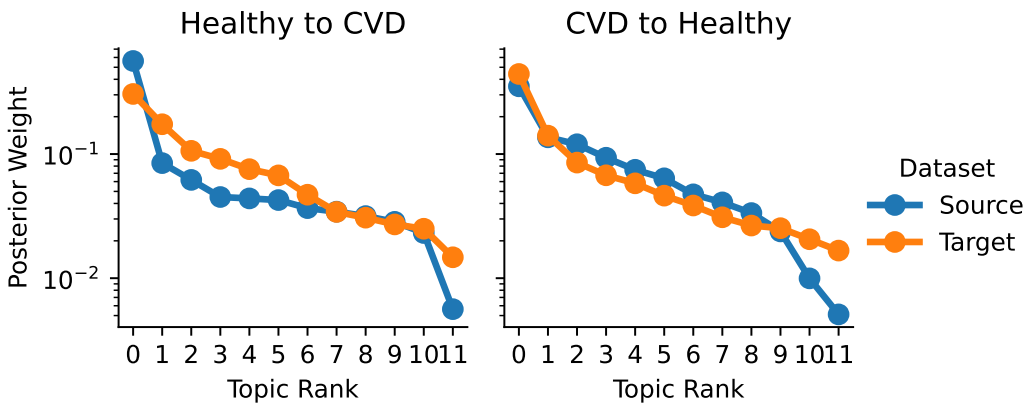


Figure C.8: Posterior topic weights $\boldsymbol{m}$ for the Adaptive Gibbs procedure in the Cleveland OOD scenario when fitting on the source dataset (blue) and when fitting on the target dataset (orange) while using the transfer prior and transferring from healthy to CVD (left) or vice-versa (right).

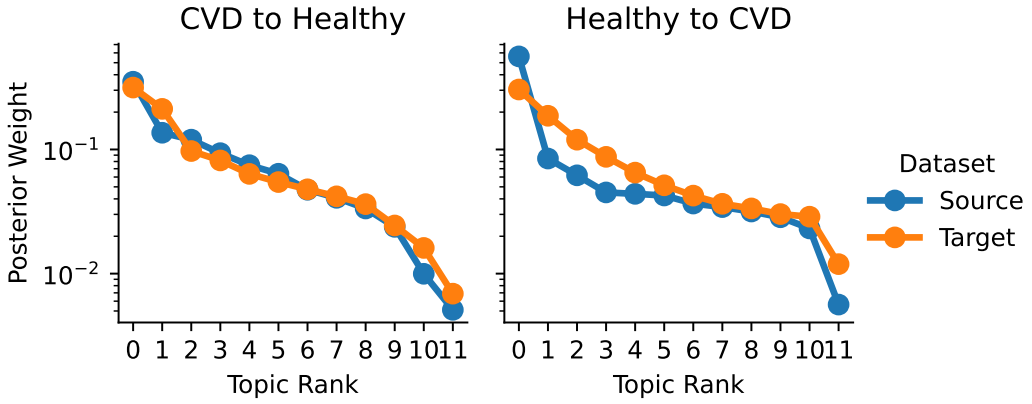


Figure C.9: Posterior topic weights $\boldsymbol{m}$ for the fixed $\lambda = 0.3$ transfer in the Cleveland OOD scenario when fitting on the source dataset (blue) and when fitting on the target dataset (orange) while using the transfer prior and transferring from healthy to CVD (left) or vice-versa (right).