PLSI: The True Fisher Kernel and beyond

IID Processes, Information Matrix and Model Identification in PLSI*

Jean-Cédric Chappelier and Emmanuel Eckard

School of Computer and Communication Sciences
École Polytechnique Fédérale de Lausanne, Switzerland

Abstract. The Probabilistic Latent Semantic Indexing model, introduced by T. Hofmann (1999), has engendered applications in numerous fields, notably document classification and information retrieval. In this context, the Fisher kernel was found to be an appropriate document similarity measure. However, the kernels published so far contain unjustified features, some of which hinder their performances. Furthermore, PLSI is not generative for unknown documents, a shortcoming usually remedied by "folding them in" the PLSI parameter space.

This paper contributes on both points by (1) introducing a new, rigorous development of the Fisher kernel for PLSI, addressing the role of the Fisher Information Matrix, and uncovering its relation to the kernels proposed so far; and (2) proposing a novel and theoretically sound document similarity, which avoids the problem of "folding in" unknown documents. For both aspects, experimental results are provided on several information retrieval evaluation sets.

1 Introduction

Ten years ago, the "Probabilistic Latent Semantic Indexing" (PLSI) model [10,11,12] opened the road to the representation of documents as mixture proportions of so-called "latent topics". This model proved useful and led to several applications on textual data [6,14,18,26,27], audio data [1] and images [5,16,19,20,24].

In this context, the cosine similarity, originally used without much theoretical justification to evaluate the semantic similarities between documents, was replaced by the Fisher kernel similarity [11], which has a better theoretical basis. However, the kernels published so far contain unjustified features, some of which hinder their performances. The first contribution of this paper, detailed in Sect. 3.1 consists in the introduction of a new, rigorous development of this Fisher kernel, that uncovers its canonical form, and shows how it relates to the kernel originally proposed by Hofmann [11].

* Work supported by projects 200021-111817 and 200020-119745 of the Swiss National Science Foundation.

W. Buntine et al. (Eds.): ECML PKDD 2009, Part I, LNAI 5781, pp. 195–210, 2009.

Moreover, one major shortcoming of PLSI comes from its tendency to overfit [2,4,23] and its non-generative nature for new document models.[1] In the context of information retrieval, this is usually remedied by "folding in" the queries into the PLSI parameter space [9,10]. A number of extensions and alternatives have been proposed to address these issues: latent Dirichlet allocation [4], undirected PLSI [28], correlated topic models [3], rate adapting Poisson models [7]; but they come at the price of an increased complexity, especially regarding the runtime cost for the learning algorithms. The second contribution of this paper, detailed in Sect. 4, targets the "folding-in" phase for queries by introducing a new, theoretically grounded, document–query similarity that entirely avoids it. This novel approach is compared to Fisher kernel similarities for PLSI.

Finally, the third contribution of this paper, detailed in Sect. 5, lies in new experimental results on a large collection coming from the TREC–AP evaluation corpus. Up to authors' knowledge, it is the first time that PLSI is evaluated on an IR corpus of over 7000 documents and one million word occurrences.

2 PLSI Document Model and Similarity

PLSI is a latent topic-based model for textual document classification and Information Retrieval [10,12]. Documents are modeled as occurrences of successive random choices of document–term couples (d, w), knowing some topic $z \in Z$: iteratively, a topic z is chosen with probability $P(z)$; a term w and a document model d are then chosen, with probabilities $P(w|z)$ and $P(d|z)$ respectively. In PLSI, w and d are assumed to be independent knowing z; the probability of occurrence of a pair (d, w) thus being

$$P(d,w) = \sum_{z \in Z} P(z)\, P(w|z)\, P(d|z) \ . \tag{1}$$

A document realization $\widehat{d_0}$ from a collection C is modeled as the set of all (d, w) pairs sharing the same document model d_0: $\widehat{d_0} = \{(d_0, w) \in C\}$. As done in the PLSI literature to simplify expressions, we will henceforth not distinguish between document model d and its concrete realization $\widehat{d}$ in the collection.

The parameters of PLSI are $\theta = \big\{P(z), P(w|z), P(d|z)\big\}$, for all possible z, w and d in the model.[2] These parameters can be estimated over a document collection through (tempered) Expectation-Maximization (EM) [10,12].

Regarding document similarity measures for PLSI, Fisher kernels yielded significant improvements in performance over the original formulations that used the usual cosine measure [11].

[1] Notice, however, that PLSI is indeed a generative model for new occurrences of *already known* document models.

[2] Other equivalent parameterizations are also possible; for instance using $P(d)$, $P(w|z)$ and $P(z|d)$.

However, the original Fisher kernel for PLSI derived by Hofmann [11] was later found to neglect the contribution of the Fisher information matrix $G(\theta)$,[3] and to contain a normalization by document length $|d|$. Variants of the Fisher kernel were thus introduced [21]:

- the observation about renormalization by document length yielded the development of a Fisher kernel not normalized by $|d|$;
- from the observation about the information matrix stemmed the development of a "*DFIM*" kernel (for "Diagonal Fisher Information Matrix"), which takes the diagonal components of $G(\theta)$ into account. By contrast, whenever required, we shall name "*IFIM*" (Identity Fisher Information Matrix) the kernels that do not.

3 IID Processes Perspective on PLSI Fisher Kernels

3.1 IID Derivation of the Fisher Kernel

The Fisher kernel provides a similarity measure among instances of probabilistic models [13]: for two instances X and Y of a given family of stochastic models $P(X|\theta)$ parameterized with θ, it is defined as

$$K(X,Y) = U_X(\theta)^T G(\theta)^{-1} U_Y(\theta) \ ,$$

where $U_X(\theta)$ is the gradient of the log-likelihood: $U_X(\theta) = \nabla_\theta \log P(X|\theta)$, and the Fisher information matrix $G(\theta)$ is the covariance of $U_X(\theta)$:

$$G(\theta) = \mathbf{E}_X[U_X(\theta)\ {U_X}^T(\theta)] \ .$$

Lemma 1. *The Fisher kernel between two instances X_1^n and Y_1^m of an independent and identically-distributed (i.i.d.) stochastic process is the sum of the Fisher kernels between the individual random variables X_i and Y_j, divided by the number of variables in the processes:*

$$K(X_1^n, Y_1^m) = \frac{1}{n}\frac{1}{m}\sum_{i=1}^{n}\sum_{j=1}^{m} K(X_i, Y_j) \ .$$

We refer to Appendix A for the proof.

The Fisher kernel for PLSI can be derived in a manner which stems directly from the definition of the PLSI model as an i.i.d. process of (d,w) pairs probabilized by (1).

In this context, $X_i = (d,w)$ and $Y_i = (q,w')$ (for two document models d and q, and terms w and w' from vocabulary V). Through Lemma 1, the Fisher kernel for PLSI as an i.i.d. process is:

$$K^{\mathrm{IID}}(d,q) = \frac{1}{|d|}\frac{1}{|q|}\sum_{w\in V}\sum_{w'\in V} n(d,w)\, n(q,w')\, K\Big((d,w),(q,w')\Big) \ , \qquad (2)$$

[3] Hofmann identified $G(\theta)$ with the identity matrix through a reparametrization suited for multinomial models; however, PLSI is neither a multinomial, nor in an exponential family, and $G(\theta)$ may significantly differ from identity in such cases.

where $n(d,w)$ is the number of occurrences of word w in query d, $|d| = \sum_w n(d,w)$ is the length of document d, and $K\big((d,w),(q,w')\big)$ is the "atomic kernel" between a pair of terms w and w' belonging to documents d and q respectively. This "atomic kernel" can be written as (see Appendix B for details): $K\left((d,w),(q,w')\right) =$

$$\frac{P(d|z)}{P(d,w)}\frac{P(q|z)}{P(q,w')} \cdot \sum_z \left[P(w|z)P(w'|z)\alpha(z) + \delta_{w,w'}\gamma(w,z) \right] \ , \tag{3}$$

with $\delta_{w,w'} = 1$ if $w = w'$ and 0 otherwise, and

$$\alpha(z) = \begin{cases} P(z) & \text{in IFIM,} \\ \left(\sum_{d \in C} \sum_{w \in V} n(d,w) \left(\frac{P(w|z)P(d|z)}{P(d,w)} \right)^2 \right)^{-1} & \text{in DFIM,} \end{cases}$$

$$\text{and } \gamma(w,z) = \begin{cases} P(w|z)P^2(z) & \text{in IFIM,} \\ \left(\sum_{d \in C} n(d,w) \left(\frac{P(d|z)}{P(d,w)} \right)^2 \right)^{-1} & \text{in DFIM.} \end{cases}$$

Injecting (3) into (2) leads to

$$K^{\mathrm{IID}}(d,q) = \sum_{w \in V} \sum_{w' \in V} \widehat{P}(w|d) \cdot \widehat{P}(w'|q) \cdot \frac{P(d|z)}{P(d,w)} \frac{P(q|z)}{P(q,w)} \cdot \cdot \sum_z \left[P(w|z)P(w'|z) \cdot \alpha(z) + \delta_{w,w'}\gamma(w,z) \right] \ , \tag{4}$$

where $\widehat{P}(w|d) = \frac{n(d,w)}{|d|}$.

3.2 Relation between $K^{\mathrm{IID}}(d,q)$ and $K^{\mathrm{H}}(d,q)$

Hofmann's original development of the Fisher kernel [11], $K^H(d,q) =$

$$\sum_{z \in Z} \frac{P(z|d)P(z|q)}{P(z)} + \sum_{w \in V} \widehat{P}(w|d)\widehat{P}(w|q) \sum_{z \in Z} \frac{P(z|d,w)P(z|q,w)}{P(w|z)} \ ,$$

uses the assumption that

$$\sum_{w \in V} \frac{\widehat{P}(w|d)}{P(w|d)} P(w|z) \simeq 1 \ . \tag{5}$$

Introducing

$$\zeta(d,z) = \sum_{w \in V} n(d,w) \frac{P(w|z)}{P(d,w)} \ , \tag{6}$$

Equation (4) becomes:

$$K^{\mathrm{IID}}(d,q) = \frac{1}{|d|}\frac{1}{|q|} \sum_{z \in Z} P(d|z)P(q|z) \cdot \cdot \left[\alpha(z)\zeta(d,z)\zeta(q,z) + \sum_{w \in V} \frac{n(d,w)}{P(d,w)} \frac{n(q,w)}{P(q,w)} \gamma(w,z) \right] \ . \tag{7}$$

Injecting (5) into (6) entails that $\zeta(d,z) \simeq \frac{|d|}{P(d)}$. Furthermore, noticing that

$$P(d) \simeq \frac{|d|}{|C|}, \qquad \text{where } |C| = \sum_{w \in V} \sum_{d \in C} n(d,w) \ , \tag{8}$$

is experimentally verified to a precision inferior to 1%, we can state that $\zeta(d,z) \simeq |C|$. In the IFIM case, this leads to

$$K^{\mathrm{IID}}(d,q) \simeq K^{\mathrm{H}}(d,q) \ .$$

Thus, at the price of assumptions (5) and (8), the Hofmann IFIM kernel can be seen as an approximation of the IID one.

In the DFIM case, the equivalence is as not exact, as the DFIM α and γ are not the same in the IID kernel and Hofmann's:

	Hofmann	IID
$\alpha^{-1}(z)$	$\sum_d \frac{P(z\|d)^2}{P(z)}$	$\sum_d \sum_w n(d,w) \left(\frac{P(w\|z)P(d\|z)}{P(d,w)}\right)^2$
$\gamma^{-1}(w,z)$	$\sum_d \widehat{P}^2(w\|d) \left(\frac{P(d\|z)}{P(d,w)}\right)^2$	$\sum_d n(d,w) \left(\frac{P(d\|z)}{P(d,w)}\right)^2$

3.3 Implementation of the IID Kernel

In practice, (7) is more efficiently computed by taking into account that $\alpha(z)$ depends only on z (and neither on d nor on w): $\alpha(z)$ and $\zeta(d,z)$ can be precomputed once for all for the entire corpus. $\gamma(w,z)$ and $\zeta(q,z)$ are best computed for each query, as to take advantage of the limited number of different terms present in a query: $\gamma(w,z)$ is computed only for $w \in q$ (i.e. $w \in V$ such as $n(w,q) > 0$). Such processing is an order of magnitude quicker using these precomputations.

Furthermore, the computation of all Fisher kernels can be decomposed into two independent parts K_z and K_w which stem from the contributions of the latent categories and of the terms, respectively; for instance using (7): $K^{\mathrm{IID}}(d,q) = K_z^{\mathrm{IID}}(d,q) + K_w^{\mathrm{IID}}(d,q)$, where

$$K_z^{\mathrm{IID}}(d,q) = \frac{1}{|d|}\frac{1}{|q|} \sum_{z \in Z} P(d|z)P(q|z)\,\alpha(z)\,\zeta(d,z)\,\zeta(q,z) \ ,$$

and

$$K_w^{\mathrm{IID}}(d,q) = \frac{1}{|d|}\frac{1}{|q|} \sum_{z \in Z} \left[P(d|z)P(q|z) \cdot \sum_{w \in V} \frac{n(d,w)}{P(d,w)} \frac{n(q,w)}{P(q,w)} \gamma(w,z) \right] \ .$$

4 Avoiding the Folding-In of Queries

The second contribution of this paper consists in the development of a document similarity measure for PLSI that entirely removes the so-called "folding-in" phase

for unknown document models q (typically the queries in Information Retrieval), and all the problems related to the learning of new parameters $P(q|z)$, as well as their adequacy with already existing ones: changing from $P(q|z) = 0$ to $P(q|z) > 0$ should imply the rescaling of all the $P(d|z)$ according to $\sum_\delta P(\delta|z) = 1$, i.e. by a factor $1/(1 + P(q|z)) \simeq 1 - P(q|z)$.

To remove folding-in entirely, we follow the Language-Model approach [22,29] and consider queries, not as new document models for which new parameters $P(q|z)$ must be learned, but rather as new occurrences of already learned document models.

The retrieval problem thus simply turns into model identification (rather than learning of new model): for a given query q, which are the (already learned) models d best representative of q?

One usual way to address such a question is to minimize the Kullback-Leibler divergence between the empirical distribution (q) and the model distribution (d) [15]:

$$\mathcal{S}_{\mathrm{KL}}(d,q) = -\mathrm{KL}\left(\widehat{P}(w|q), P(w|d)\right) = \sum_{w \in q \cap d} \widehat{P}(w|q) \log \frac{P(w|d)}{\widehat{P}(w|q)} \ , \qquad (9)$$

where $w \in q \cap d$ denotes all the words appearing in q (i.e. $n(q,w) > 0$) such that $P(d,w) > 0$.

Notice that this formulation uses $\widehat{P}(w|q) = n(q,w)/|q|$, for which no learning is required, as opposed to $P(w|q)$, for which unknown parameters $P(q|z)$ have to be estimated (usually done through "folding in").

5 Experiments

We are thus faced with 13 different document similarity measures: $\mathcal{S}_{\mathrm{KL}}(d,q)$, which does not require query folding-in (Sect. 4), and 12 Fisher kernel variants: the two IFIM models K^{H} and K^{IID}, and the corresponding DFIM kernels $K^{\text{DFIM-H}}$ and $K^{\text{DFIM-IID}}$; as well as, separately, the K_z and K_w components of all these kernels (as defined in Sect. 3.3). Across $K^{(\text{DFIM-})\mathrm{H}}$ and $K^{(\text{DFIM-})\mathrm{IID}}$, the latter provide 8 different kernels.

The following questions arise regarding these 13 kernels:

1. Are there significant differences between all the possible variants of the Fisher kernel and which one is the best?
2. Can the new approaches here proposed compete with the Hofmann kernel variant?
3. How do these compare to the IR state-of the art model, BM25 [25], especially on a large document collection?

To address these questions in line with previously published work on PLSI, we experimented on the standard Information Retrieval benchmarks from the SMART collection[4]: CACM, CISI, MED, CRAN and TIME. We furthermore

[4] `ftp://ftp.cs.cornell.edu/pub/smart/`

Table 1. Characteristics of the document collections used for evaluation

	CACM	CRAN	TIME	CISI	MED	AP89_01XX
# Terms (stems)	4 911	4 063	13 367	5 545	7 688	13 379
# occurrences ($\|C\|$)	90 927	120 973	114 850	87 067	76 571	1 321 482
Documents						
#	1 587	1 398	425	1 460	1 033	7 466
avg. $\|d\|$	56.8	85.1	268.6	56.7	73.8	177.2
Queries						
#	64	225	83	112	30	50
avg. $\|q\|$	12.7	8.9	8.2	37.7	11.4	79.3

explored the limits of PLSI learning tractability, and experimented on a significantly bigger corpus[5] consisting of a subpart of the TREC–AP 89 corpus [8]. For tractability reasons, we kept only the 7466 first documents of this collection,[6] and queries 1 to 50. The main characteristics of the evaluation corpora are given in Table 1.

For experiments on the SMART collection, 6 runs with different learning initial conditions were performed for all the models, and for different numbers of topics: $|Z| \in \{1, 2, 8, 16, 32, 64, 128\}$, totalling 2730 experiments. For the TREC-AP part, due to its size, a single run was performed for each $|Z| \in \{1, 32, 48, 64, 80, 128\}$, totaling 78 experiments.

For all the experiments, stemming was performed using the Porter algorithm of Xapian.[7] Evaluation results were obtained using the standard `trec_eval` tool.[8] We here use the standard Mean Average Precision (MAP) to display the results: we plot the MAP against the number of latent topics $|Z|$, averaged over all experiments and with error bars corresponding to 1-standard deviation. The conclusions are exactly the same using either 5-point precision or R-precision, except on MED; we come back to this latter point at the end of this section.

The main results out of these experiments, summarized in Table 2, are:

1. The kernels K^{IID} and K^{H} have very similar performances, which experimentally validates the theoretical result that K^{H} approximates K^{IID} (see Fig. 1). $K^{\text{DFIM-IID}}$ and $K^{\text{DFIM-H}}$ are a bit less similar, due to the slightly different form that $G(\theta)$ takes. However, K^{IID} is more computationally demanding than K^{H}, at least one order of magnitude slower.

[5] Over 5 times as many documents and 10 times as many word occurrences as in the SMART collection.

[6] Documents AP890101-0001 to AP890131-0311. The EM learning for e.g. $|Z| = 128$ took 45 hours of CPU time and used 6.7 Gb of RAM on a dedicated computer server with one octo core 2-GHz Intel Xenon processor and 32 Gb of memory.

[7] `http://xapian.org/`

[8] `http://trec.nist.gov/trec_eval/`

Table 2. Main results and conclusions out of 2808 experiments over 13 models on 6 corpora

		CACM	CRAN	TIME	CISI	MED	AP89_01XX
Results	BM25 MAP	**31.4**	**42.4**	**69.2**	12.3	52.3	19.7
	Best PLSI model MAP	30.0	39.6	60.8	**20.2**	**53.8**	**21.6**
	Best PLSI model is:	$K_w^{\text{DFIM-H}}$	$\mathcal{S}_{\text{KL}}$	$K_w^{\text{DFIM-H}}$	K_w^{H}	K^{H}	$K_w^{\text{DFIM-H}}$
	for $\lvert Z\rvert =$	32	128	8	8	32	48
	K_w^{H} MAP	**30.0**	33.6	55.6	**20.2**	49.8	16.5
	$K_w^{\text{DFIM-H}}$ MAP	23.2	37.0	**60.8**	15.6	45.5	**21.6**
	$\mathcal{S}_{\text{KL}-128}$ MAP	22.9	**39.6**	49.1	19.5	**52.8**	11.4
Concl.	PLSI > BM25?	No	No	No	**YES**	**yes**	**yes**
	$\mathcal{S}_{\text{KL}-128}$ w.r.t. Fisher kernels	$<$	$\boxed{>}$	$<$	$\simeq$	$\simeq$	$<$
	DFIM $G(\theta)$ helps? (on K_w)	Yes	**Yes**	**Yes**	No	No	**Yes**

2. We can confirm that the DFIM Fisher kernels for PLSI outperform by their original IFIM versions (Fig. 1). They are furthermore dominated by their K_w component.
3. K_z deteriorates performances in general: used alone, it performs poorly; furthermore the performances of K^{H} decrease as the role of K_z becomes more important for growing $|Z|$: starting from K_w at low $|Z|$, the performances of K^{H} reach down K_z at higher $|Z|$(Fig. 3).
   On the other hand, K_w alone is always good, if not the best (Figs. 2 and 3).
4. $K_w^{\text{DFIM-H}}$ and $K^{\text{DFIM-H}}$ have similar behaviors (Fig. 3). The reason is that the normalizing role of $G(\theta)$ makes $K_z \ll K_w$ for DFIM kernels.
5. $\mathcal{S}_{\text{KL}}$ has a growing performance with $|Z|$, the number of latent-topics (Fig. 2). We had to stop at $|Z| = 128$ for tractability reasons (for both learning and evaluation running times).
6. $\mathcal{S}_{\text{KL}}$ can outperform the best Fisher kernel on CRAN and reaches similar performances on MED and CISI (Fig. 2). It should however be emphasized that the former does not require any folding-in phase for queries as the latter does.
7. The best PLSI-based kernels can perform better than the state-of-the-art BM25 model, especially on corpora which could be considered semantically more difficult: CISI, where few words are shared between queries and documents (thus particularly suited to test the extend to which a retrieval models is robust to synonymy, or "topics".[9]), MED (specialized vocabulary), and TREC–AP.

The only collection where conclusions should be more nuanced is MED: the different models do not behave in the same way at different recall values (Fig. 4); some are better at low recall and others are better at higher recall. Global measures as MAP or R-Prec cannot represent such nuances.

[9] CISI is remarkable in the sense that some query-document matches are expected between queries and documents that do not share any significant term.

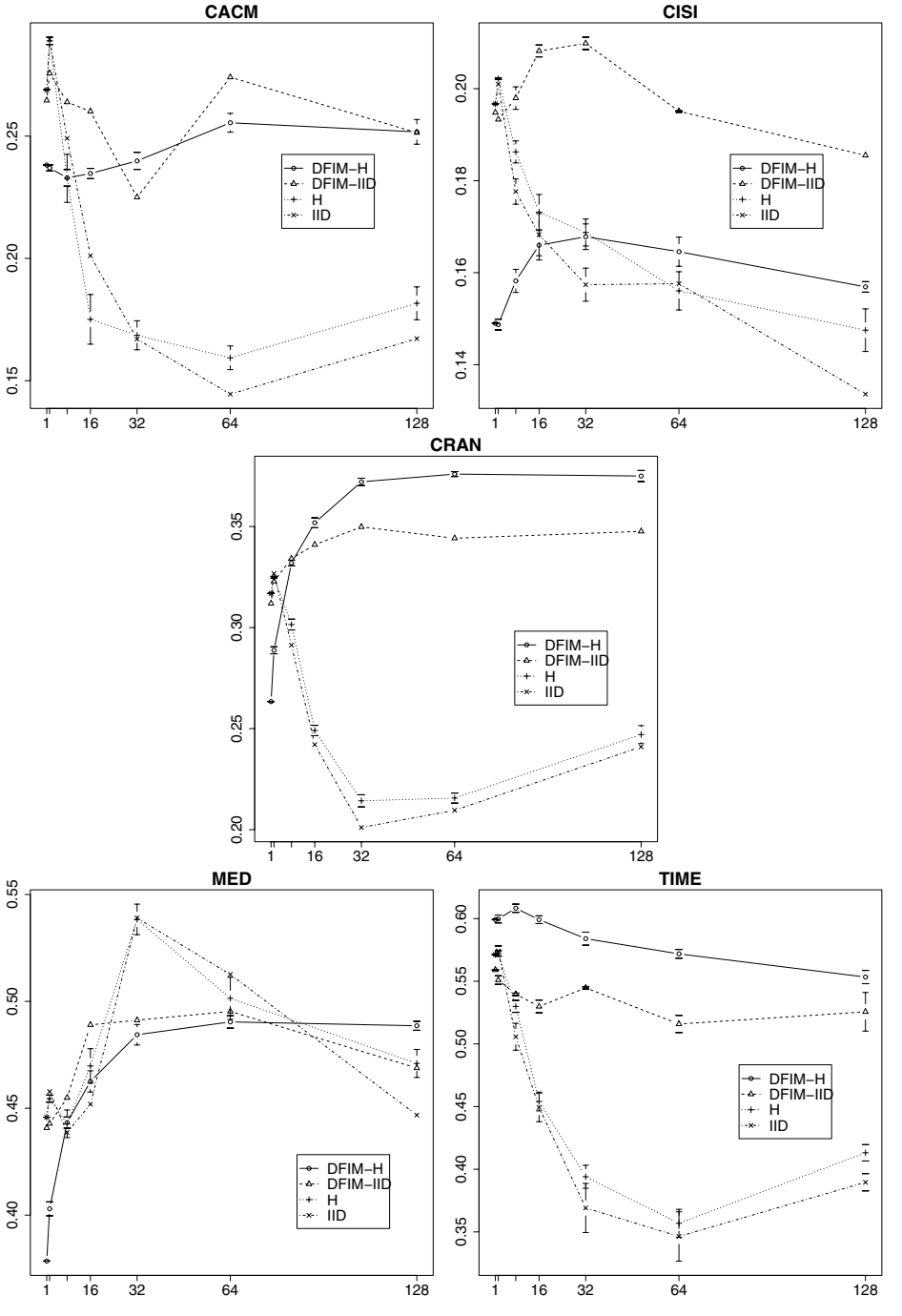

Fig. 1. Behaviors (MAP vs $|Z|$) of the K^{IID} and K^{H} kernels, both for IFIM and DFIM variants

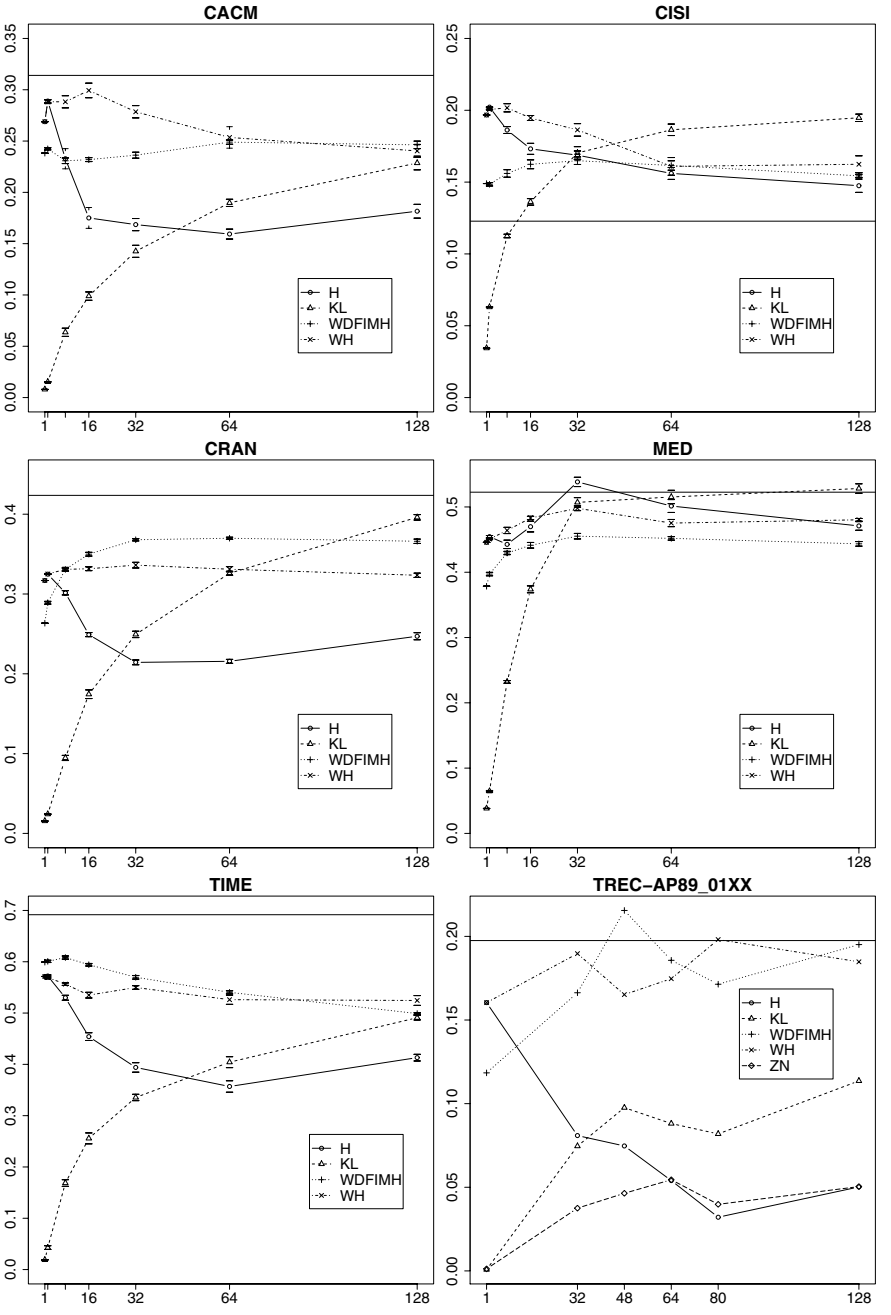

Fig. 2. Results (MAP vs $|Z|$) obtained on the six corpora considered for different models: K^{H} (H), $\mathcal{S}_{\mathrm{KL}}$ (KL), $K_w^{\text{DFIM-H}}$ (WDFIMH), and K_w^{H} (WH). The horizontal bar represents the MAP of state-of-the-art BM25 model (which does not depend on $|Z|$).

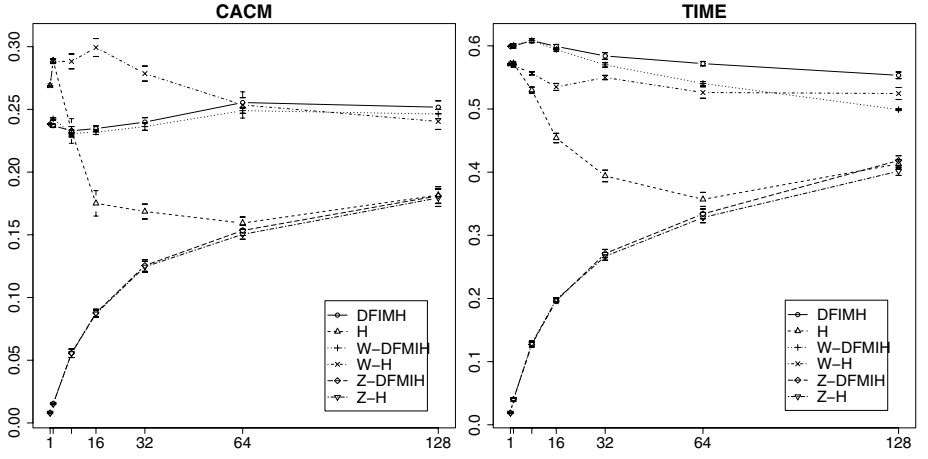

Fig. 3. Two typical examples comparing different variants of the Fisher kernel for PLSI (MAP vs $|Z|$): $K^{\text{DFIM-H}}$ (DFIMH), K^{H} (H), $K_w^{\text{DFIM-H}}$ (W-DFIMH), K_w^{H} (W-H), $K_z^{\text{DFIM-H}}$ (Z-DFIMH), and K_z^{H} (Z-H). This illustrates that the latent-topic part K_z performs poorly in comparison to the word part K_w, and impairs the combined kernels: notice how K^{H} starts from K_w^{H} at low $|Z|$, and degrades to K_z^{H} as $|Z|$ grows.

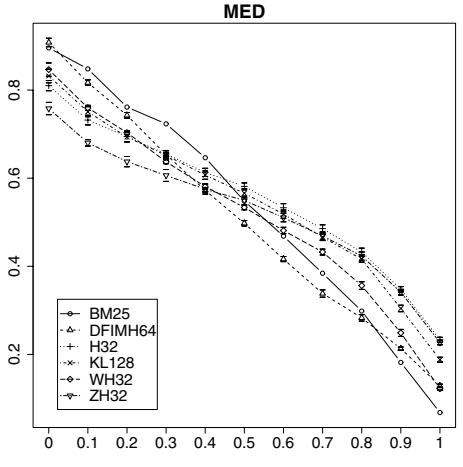

Fig. 4. Precision vs Recall curves on MED for BM25, $K^{\text{DFIM-H}}$ for $|Z|$=64 (DFIMH64), K^{H} for $|Z|$=32 (H32), $\mathcal{S}_{\text{KL}}$ for $|Z|$=128 (KL128), K_w^{H} for $|Z|$=32 (WH32), and K_z^{H} for $|Z|$=32 (YH32)

6 Conclusion

This paper offers two contributions: the first one is a rigorous development of the Fisher kernel for PLSI models, by which we uncover the kernel that properly takes into account the i.i.d. nature of PLSI, thereby explaining the underlying

reasons for normalization by document and query length in Hofmann's kernel; and restore the contribution of the Fisher Information Matrix into these kernels. The second contribution is a theoretically grounded similarity for PLSI which entirely avoids query folding-in. Furthermore, we were able to perform an Information Retrieval evaluation of PLSI on a collection much larger than the SMART collections on which it is usually evaluated.

Regarding the questions we wanted to address experimentally, we can conclude:

1. There are significant differences between all the possible variants of Fisher kernel for PLSI and the best variant is globally $K_w^{\text{DFIM-H}}$.
   Regarding the role of the Fisher information matrix, its normalizing impact improves the results on bigger collections (TIME, CRAN, TREC–AP89).
   Regarding the role of topics and terms components, K_z should clearly be neglected in favor of K_w; at least for tractable numbers of topics (small $|Z|$).
   The rigorous IID kernels $K_w^{\text{(DFIM-)IID}}$ offer performances similar to those of $K_w^{\text{(DFIM-)H}}$; this is to be expected, as the Hofmann's kernel family turns out to be an approximation of the IID one. Since the kernels $K_w^{\text{(DFIM-)IID}}$ are computationally more expensive, $K_w^{\text{(DFIM-)H}}$ should be preferred in practice.
2. The new approach which avoids query folding-in can compete with the best Fisher kernel variants, especially for high number of topics.
3. These models (either KL divergence or Fisher kernels) can compete with BM25, especially on corpora which could be considered semantically more difficult as CISI, MED and TREC–AP.

We thus experimentally confirm that topic-based models as PLSI could be interesting for information retrieval in not too large but semantically difficult document collections, where documents and queries do not necessarily share at lot of terms. In such cases, we would recommend $K_w^{\text{DFIM-H}}$ as similarity measure, or $\mathcal{S}_{\text{KL}}$, (9), when sufficiently large numbers of latent topics are tractable. The advantage of the latter is the lack of folding-in phase for queries.

The overall conclusion, however, is that PLSI is not well-suited for large scale ad hoc IR, mainly because it is not a fully generative model. This model simply does not scale with the number of documents. Furthermore, as sophisticated as it can be, PLSI hardly outperforms the state-of-the-art BM25 model, at a complexity price that is not worth paying. PLSI might be interesting and better than state-of-the-art models for large numbers of latent-topics, but the former limitation (huge number of parameters) makes such levels intractable in practice. It may indeed be the case, especially for bigger document collections, that a *much* larger $|Z|$ improves the performances of K_z or $\mathcal{S}_{\text{KL}}$, but, due to the non-scalability of PLSI, such levels won't finish training in practice, especially for the bigger document collections where they would be interesting.

References

1. Ahrendt, P., Goutte, C., Larsen, J.: Co-occurrence models in music genre classification. In: IEEE Int. Workshop on Machine Learning for Signal Processing (2005)

2. Bast, H., Weber, I.: Insights from viewing ranked retrieval as rank aggregation. In: Proc. of Int. Workshop on Challenges in Web Information Retrieval and Integration (WIRI 2005), pp. 232–239 (2005)
3. Blei, D., Lafferty, J.: A correlated topic model of *Science*. Annals of Applied Statistics 1(1), 17–35 (2007)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. Journal of Machine Learning Research 3, 993–1022 (2003)
5. Bosch, A., Zisserman, A., Munoz, X.: Scene classification via plsa. In: Proc. of the European Conf. on Computer Vision (2006)
6. Gaussier, E., Goutte, C., Popat, K., Chen, F.: A hierarchical model for clustering and categorising documents. In: Proc. of 24th BCS-IRSG Europ. Coll. on IR Research, pp. 229–247 (2002)
7. Gehler, P.V., Holub, A.D., Welling, M.: The rate adapting Poisson model for information retrieval and object recognition. In: Proc. 23rd Int. Conf. on Machine Learning, pp. 337–344 (2006)
8. Harman, D.: Overview of the fourth Text REtrieval Conference (TREC–4). In: Proc. of the 4th Text REtrieval Conf., pp. 1–23 (1995)
9. Hinneburg, A., Gabriel, H.-H., Gohr, A.: Bayesian folding-in with Dirichlet kernels for PLSI. In: Proc. of the 7th IEEE Int. Conf. on Data Mining, pp. 499–504 (2007)
10. Hofmann, T.: Probabilistic latent semantic indexing. In: Proc. of 22nd Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, pp. 50–57 (1999)
11. Hofmann, T.: Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In: Advances in Neural Information Processing Systems, vol. 12, pp. 914–920 (2000)
12. Hofmann, T.: Unsupervised learning by probabilistic latent semantic analysis. Machine Learning 42(1), 177–196 (2001)
13. Jaakkola, T., Haussler, D.: Exploiting generative models in discriminative classifiers. In: Advances in Neural Information Processing Systems, vol. 11, pp. 487–493. MIT Press, Cambridge (1999)
14. Jin, X., Zhou, Y., Mobasher, B.: Web usage mining based on probabilistic latent semantic analysis. In: Proc. of 10th Int. Conf. on Knowledge Discovery and Data Mining, pp. 197–205 (2004)
15. Lafferty, J., Zhai, C.: Document language models, query models, and risk minimization for information retrieval. In: Proc. ACM SIGIR Conf. on Research and Development in Information Retrieval (2001)
16. Lienhart, R., Slaney, M.: Plsa on large-scale image databases. In: Proc. of the 2007 Int. Conf. on Acoustics, Speech and Signal Processing, IEEE (ICASSP 2007), vol. 4, pp. 1217–1220 (2007)
17. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley, Chichester (2000)
18. Mei, Q., Zhai, C.: A mixture model for contextual text mining. In: Proc. of 12th Int. Conf. on Knowledge Discovery and Data Mining, pp. 649–655 (2006)
19. Monay, F., Gatica-Perez, D.: Plsa-based image auto-annotation: Constraining the latent space. In: Proc. ACM Int. Conf. on Multimedia, ACM MM (2004)
20. Monay, F., Gatica-Perez, D.: Modeling semantic aspects for cross-media image indexing. IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (2007)
21. Nyffenegger, M., Chappelier, J.-C., Gaussier, E.: Revisiting Fisher kernels for document similarities. In: Proc. of 17th European Conf. on Machine Learning, pp. 727–734 (2006)

22. Ponte, J.M., Croft, W.B.: A language modeling approach to information retrieval. In: 21st SIGIR Conf. on Research and Development in Information Retrieval, pp. 275–281 (1998)
23. Popescul, A., Ungar, L.H., Pennock, D.M., Lawrence, S.: Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In: Proc. of the 17th Conf. in Uncertainty in Artificial Intelligence, pp. 437–444 (2001)
24. Quelhas, P., Monay, F., Odobez, J.-M., Gatica-Perez, D., Tuytelaars, T., Gool, L.V.: Modeling scenes with local descriptors and latent aspects. In: Proc. of ICCV 2005, vol. 1, pp. 883–890 (2005)
25. Robertson, S.E., Walker, S., Jones, S., Hancock-Beaulieu, M., Gatford, M.: Okapi at TREC–3. In: Proc. of the 3rd Text REtrieval Conf. (1994)
26. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.: Probabilistic author-topic models for information discovery. In: Proc. 10th Int. Conf. on Knowl. Discovery and Data Mining, pp. 306–315 (2004)
27. Vinokourov, A., Girolami, M.: A probabilistic framework for the hierarchic organisation and classification of document collections. Journal of Intelligent Information Systems 18(2/3), 153–172 (2002)
28. Welling, M., Rosen-Zvi, M., Hinton, G.: Exponential family harmoniums with an application to information retrieval. In: Advances in Neural Information Processing Systems, vol. 17, pp. 1481–1488 (2005)
29. Zhai, C.: Statistical language models for information retrieval a critical review. Found. Trends Inf. Retr. 2(3), 137–213 (2008)

A Proof of Lemma 1

Let us consider X_1^n a n-element long instance of an i.i.d. stochastic process. To simplify notations, we shall henceforth write $X = X_1^n$. Its log-likelihood is expressed by $\log P(X) = \sum_{i=1}^n \log P(X_i)$. Linearity of derivation operators entails that the corresponding Fisher score is written $U_\theta(X) = \sum_{i=1}^n U_\theta(X_i)$.

The Fisher information matrix for n-event instances of this i.i.d. stochastic process is defined as $G_X(\theta) = E_X\left[U_\theta(X)\,U_\theta(X)^T\right]$; and for a single instance: $G_1(\theta) = E_{X_i}\left[U_\theta(X_i)\,U_\theta(X_i)^T\right]$ (which is independent of X_i since the process is i.i.d.). It can be written as:

$$\begin{aligned}
G_X(\theta) &= E_X\left[\left(\sum_{i=1}^{n} U_\theta(X_i)\right)\left(\sum_{i=j}^{m} U_\theta(X_j)\right)^T\right] \\
&= \sum_{i=1}^{n}\left(E_X\left[U_\theta(X_i)U_\theta(X_i)^T\right] + \sum_{j\neq i} E_X\left[U_\theta(X_i)U_\theta(X_j)^T\right]\right) \\
&= \sum_{i=1}^{n}\left(G_{X_i}(\theta) + E_{X_i}\left[U_\theta(X_i)\right]\sum_{j\neq i}\underbrace{E_{X_j}\left[U_\theta(X_j)\right]^T}_{=0}\right) = \sum_{i=1}^{n} G_1(\theta) \\
&= n\cdot G_1(\theta)\ .
\end{aligned}$$

The "natural gradient" ϕ_X is obtained from the ordinary gradient U_X via $\phi_X = {G_X}^{-1} U_X$ [13]. In the case of i.i.d. stochastic processes, the previous results lead to

$$\phi_X = {G_X}^{-1} U_X = \frac{1}{n} \cdot \sum_{i=1}^{n} {G_1}^{-1} U_{X_i} = \frac{1}{n} \cdot \sum_{i=1}^{n} \phi_{X_i} \ .$$

Eventually, the Fisher kernel between two instances X_1^n and Y_1^m of an i.i.d. process is given by

$$\begin{aligned} K(X_1^n, Y_1^m) = {\phi_X}^T G_1 \, \phi_Y &= \left(\frac{1}{n} \sum_{i=1}^{n} {G_1}^{-1} U_{X_i} \right)^T G_1 \left(\frac{1}{m} \sum_{i=j}^{m} {G_1}^{-1} U_{Y_j} \right) \\ &= \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{i=j}^{m} U_{X_i}^T {G_1}^{-1} U_{Y_j} = \frac{1}{n \cdot m} \sum_{i=1}^{n} \sum_{i=j}^{m} K(X_i, Y_j) \ . \end{aligned}$$

B Development of the PLSI Atomic Fisher Kernel

B.1 Fisher Score $U_{(d,w)}$

The Fisher score for PLSI is written $U_{(d,w)}(\theta) = \nabla_\theta \log P(d,w)$. Derivations are performed with respect to $P(z)$, $P(d|z)$ and $P(w|z)$ respectively for all terms w, documents d and categories z. Let us write $\tilde{d}$ and $\tilde{w}$ the indices of the document and term with respect to which derivations of $P(d,w)$ are performed. Then,

$$U_{(d,w)}(\theta) = \nabla_\theta \log P(d,w) = \begin{pmatrix} \frac{\partial \log P(w,d)}{\partial P(z)} \\ \frac{\partial \log P(w,d)}{\partial P(\tilde{w}|z)} \\ \frac{\partial \log P(w,d)}{\partial P(\tilde{d}|z)} \end{pmatrix} = \begin{pmatrix} \frac{P(w|z)P(d|z)}{P(w,d)} \\ \delta_{\tilde{w}w} \frac{P(d|z)P(z)}{P(w,d)} \\ \delta_{\tilde{d}d} \frac{P(w|z)P(z)}{P(w,d)} \end{pmatrix} \ ,$$

where $\delta_{\tilde{w}w} = 1$ if $\tilde{w} = w$ and 0 else (and similarly for d).

Note that two terms $\frac{\partial \log P(w,d_i)}{\partial P(\tilde{d}|z)}$ and $\frac{\partial \log P(w,d_j)}{\partial P(\tilde{d}|z)}$ are both non-zero if and only if $i = j$, that is if a document is compared to itself. Since this case is trivial, the terms of $U_{(d,w)}(\theta)$ that stem from the derivation w.r.t. $P(\tilde{d}|z)$ can be ignored in the kernel.

At this stage, for the sake of consistency with Hofmann's derivation, a square root reparametrization can be introduced:

$$\theta(z) \to \rho(z) = 2\sqrt{P(z)} \quad \text{and} \quad \theta(w|z) \to \rho(w|z) = 2\sqrt{P(w|z)} \ .$$

In the DFIM case, this reparametrization eventually cancels out; however, in the IFIM case, its contribution remains, making it a necessary step in the derivation. With this reparametrization, $\frac{\partial P(z)}{\partial \rho(z)} = \frac{1}{2}\rho(z) = \sqrt{P(z)}$ and $\frac{\partial P(w|z)}{\partial \rho(\tilde{w}|z)} = \delta_{\tilde{w}w} \frac{1}{2} \rho(w|z) = \delta_{\tilde{w}w} \sqrt{P(w|z)}$. Hence,

$$U_{(d,w)}(\rho) = \begin{pmatrix} \sqrt{P(z)} \, \frac{P(w|z)P(d|z)}{P(d,w)} \\ \delta_{\tilde{w}w} \sqrt{P(w|z)} \, \frac{P(d|z)P(z)}{P(w,d)} \end{pmatrix} \ .$$

B.2 Fisher Information Matrix $G(\theta)$

The Fisher information matrix is written

$$G(\theta) = E_{(d,w)}\left[U_{(d,w)}(\theta)\, {U_{(d,w)}}^T(\theta)\right] \quad .$$

Let us consider the parts of the matrix G which stem from $U_{(d,w)}(z) = \frac{\partial \log P(d,w)}{\partial P(z)}$ (noted G_z) and from $U_{(d,w)}(\tilde{w}|z) = \frac{\partial \log P(d,w)}{\partial P(\tilde{w}|z)}$ (noted G_w). The diagonal of $G(\theta)$ can be approximated by (e.g. [17]):

$$G_z(z) = \sum_{(d,w)\in C} U_{(d,w)}(z)^2, \qquad G_w(\tilde{w},z) = \sum_{(d,w)\in C} U_{(d,w)}(\tilde{w}|z)^2 \quad ,$$

which leads to

$$G_z(z) = \sum_{d\in C}\sum_{w\in d} n(d,w)\left(\frac{P(w|z)P(d|z)}{P(d,w)}\right)^2 \qquad \text{and}$$

$$G_w(w,z) = \sum_{d\in C} n(d,w)\left(\frac{P(w|z)P(z)}{P(d,w)}\right)^2 \quad .$$

B.3 Fisher Kernel

The expressions for $U_{(d,w)}$ and G, can be assembled into the Fisher kernel $K\big((d,w_d),(q,w_q)\big) = U_{(d,w_d)}\, G^{-1}\, U_{(q,w_q)}$, written as

$$\begin{aligned}
K\big((d,w_d),(q,w_q)\big) &= K_z\big((d,w_d),(q,w_q)\big) + K_w\big((d,w_d),(q,w_q)\big), \text{ where}\\
K_z\big((d,w_d),(q,w_q)\big) &= \sum_z U_{(d,w_d)}\, G_z(z)^{-1}\, U_{(q,w_q)} \quad , \text{ and}\\
K_w\big((d,w_d),(q,w_q)\big) &= \sum_z\sum_{\tilde{w}} U_{(d,w_d)}\, G_w(\tilde{w},z)^{-1}\, U_{(q,w_q)}\\
&= \delta_{w_d w_q}\sum_z U_{(d,w_d)}\, G_w(w_d,z)^{-1}\, U_{(q,w_q)} \quad .
\end{aligned}$$

Eventually, $K\Big((d,w_d),(q,w_q)\Big) =$

$$\sum_z\Big[\ \frac{P(w_d|z)P(d|z)}{P(d,w_d)}\frac{P(w_q|z)P(q|z)}{P(q,w_q)}P(z)G_z(z)^{-1} + \delta_{w_d,w_q}\frac{P(d|z)P(q|z)P^2(z)}{P(d,w_d)P(q,w_d)}P(w_d|z)G_w(w_d,z)^{-1}\Big] \quad .$$

Note the normalizing role of $G(\theta)$: all the terms not depending on d in U_d will cancel out, as they can be factorized in $G_z(z)$ and $G_w(w,z)$, respectively: $P(z)$ cancels out in K_z and $P(w|z)P^2(z)$ cancels out in K_w.