

Adaptive Distributional Extensions to DFR Ranking

Casper Petersen
University of Copenhagen
cazz@di.ku.dk

Kalervo Järvelin
University of Tampere
Kalervo.Jarvelin@staff.uta.fi

Jakob Grue Simonsen
University of Copenhagen
simonsen@di.ku.dk

Christina Lioma
University of Copenhagen
c.lioma@di.ku.dk

ABSTRACT

Divergence From Randomness (DFR) ranking models assume that informative terms are distributed in a corpus differently than non-informative terms. Different statistical models (e.g. Poisson, geometric) are used to model the distribution of non-informative terms, producing different DFR models. An informative term is then detected by measuring the divergence of its distribution from the distribution of non-informative terms. However, there is little empirical evidence that the distributions of non-informative terms used in DFR actually fit current datasets. Practically this risks providing a poor separation between informative and non-informative terms, thus compromising the discriminative power of the ranking model. We present a novel extension to DFR, which first detects the best-fitting distribution of non-informative terms in a collection, and then adapts the ranking computation to this best-fitting distribution. We call this model Adaptive Distributional Ranking (ADR) because it adapts the ranking to the statistics of the specific dataset being processed each time. Experiments on TREC data show ADR to outperform DFR models (and their extensions) and be comparable in performance to a query likelihood language model (LM).

CCS Concepts

•Information systems → Retrieval models and ranking;

1. INTRODUCTION

Early work on automatic indexing [3, 8, 9] observed that *informative* words, e.g. those belonging to a technical vocabulary, are distributed in a document collection differently than *non-informative* words, e.g. those usually treated as stopwords. Specifically, the difference in the distributions of these two different types of words was that informative words were observed to appear more densely in few so-called *elite* documents. Based on this, non-informative words were modelled by a Poisson distribution, and it was hypothesised

that an informative word can be detected by measuring the extent to which its distribution deviates from a Poisson distribution, or, in other words, by testing the hypothesis that the word distribution on the whole document collection does not fit the Poisson model [2]. This so-called 2-Poisson model has led to well-known ranking models such as BM25 [18] or Divergence from Randomness (DFR) [2, 14].

For DFR ranking models in particular (presented in Section 2), the idea that non-informative terms tend to follow a specific distribution is central. Different distributions (e.g. the Poisson or geometric distribution) can be used to obtain different DFR ranking models. The goal is to choose a distribution that provides a good fit to the empirical distribution of non-informative terms in C . However, there is little empirical evidence that the distributions used in DFR actually fit current datasets. Practically this risks providing a poor separation between informative and non-informative terms, thus compromising the discriminative power of the ranking model.

Motivated by this, we present a novel extension to DFR, which first detects the best-fitting distribution (among a candidate set of distributions) of non-informative terms in a collection, and then adapts the ranking computation to this “best-fitting” distribution. We call this model Adaptive Distributional Ranking (ADR, presented in Section 3) because it adapts the ranking to the statistics of the specific dataset being processed each time. With our approach, only one ranking model is obtained per collection: the one that provides the best fit to the distribution of non-informative terms in that collection. Evaluation on TREC datasets (Section 4) shows that our ADR models outperform the original DFR models, as well as more recent DFR extensions and perform on a par with a query likelihood LM [17].

2. DIVERGENCE FROM RANDOMNESS

Given a query q and a document d , a DFR ranking model estimates the relevance $R(q, d)$ of d to q as [2]:

$$R(q, d) = \sum_{t \in q} f_{t,q} \cdot (-\log_2 P_1) \cdot (1 - P_2) \quad (1)$$

where t is a term, $f_{t,q}$ is the frequency of t in q , $-\log_2 P_1$ is the information content of t in d , and $(1 - P_2)$ is the risk of accepting t as a descriptor of d 's topic. We explain each of these two components next.

The information content P_1 measures the divergence of $f_{t,d}$ from $f_{t,C}$, where a higher divergence means more information is carried by t in d . The assumption is that terms that bring little information are distributed over all documents in a certain way across the entire corpus; different

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

CIKM'16, October 24-28, 2016, Indianapolis, IN, USA

© 2016 ACM. ISBN 978-1-4503-4073-1/16/10...\$15.00

DOI: <http://dx.doi.org/10.1145/2983323.2983895>

distributions (called “models of randomness” in DFR) give rise to different DFR ranking models. For example, using a Poisson distribution, P_1 is:

$$P_1(t, \lambda, d) = e^{-\lambda} \lambda^{\hat{f}_{t,d}} / \hat{f}_{t,d}! \quad (2)$$

where $\lambda = f_{t,C} / |C|$ is the parameter of the Poisson distribution, and $\hat{f}_{t,d} = f_{t,d} \cdot \log_2(1 + c \cdot \text{avg_l} / |d|)$ is a logarithmic term frequency normalisation where c is a free parameter, avg_l is the average document length in the collection, and $|d|$ is the length of d (the number of terms). Eqn. 2 returns the probability of seeing $\hat{f}_{t,d}$ occurrences of t in d and effectively tests whether t 's distribution on C fits the Poisson distribution. If the probability of obtaining $\hat{f}_{t,d}$ occurrences of t is low, then t carries a high amount of information [2]. In addition to the Poisson distribution used in this example, DFR models can be instantiated also with the geometric, tf-idf (I_n), tf-idf (I_F) and tf-expected-idf (I_n^e) [2].

The second component of Eqn. 1, the information gain, P_2 , is the conditional probability of encountering $f_{t,d} + 1$ occurrences of t in d . If P_2 is high, the risk $(1 - P_2)$ associated with accepting t as a descriptor of d is low. Such “after-effect” sampling was taken [2] to be normalised either with Laplace: $P_2 = f_{t,d} / (f_{t,d} + 1)$ or Bernoulli normalisation: $P_2 = (f_{t,C} + 1) / (n_t \cdot f_{t,d} + 1)$ where n_t is the number of documents where t occurs.

3. ADAPTIVE DISTRIBUTIONAL RANKING

Our ADR models follow the basic DFR rationale presented above, but adapt the computation of P_1 in Eqn. 1 to the best-fitting distribution of non-informative term collection frequencies for each collection. The ADR algorithm is shown below.

Algorithm $ADR(C)$

Input: Collection C ; $\mathbb{Y} = \{\text{parameterised statistical models}\}$

1. **for** the set T of non-informative terms in C
2. **for** each model $\mathcal{M} \in \mathbb{Y}$
3. find \mathcal{M} 's best parameters for fitting T
4. find $\hat{\mathcal{M}}$ that fits T best
5. replace P_1 in Eqn. 1 by $\hat{\mathcal{M}}$
6. **return** an ADR model derived for C

Given some collection C and a set \mathbb{Y} of candidate parameterised statistical models, we determine for each candidate model the optimal parameter values (if any) that make it fit how non-informative terms are distributed in C (step 3 in the algorithm), select the best-fitting distribution (step 4), and plug it in the place of P_1 in Eqn. 1 (step 5). The output is an ADR model adapted specifically to C (i.e. a collection-specific ranking model). In this work, we use as candidate parameterised statistical distributions all the discrete models in [16], namely the geometric, negative Binomial, Poisson, power law and Yule–Simon. However, any other set of candidate distributions can be used in step 2. We explain steps 3 – 5 next.

3.1 Step 3. Optimal parameter fitting

A parameterised statistical distribution $\mathcal{M} = \{g(T|\theta) : \theta \in \Theta\}$ where $\theta \in \Theta$ is a real or integer-valued vector, is a family of probability density or mass functions $g(T|\theta)$. *Parameter estimation* is then the problem of finding, among all the probability density/mass functions of \mathcal{M} , the $\hat{\theta}$ that most likely generated T . We estimate $\hat{\theta}$ with maximum likelihood estimation as follows: We seek the density/mass func-

tion that makes the observed term frequencies “most likely” [13] using a likelihood function, $\mathcal{L}(\theta|T)$, that specifies the likelihood of θ given T . $\hat{\theta}$ is then obtained by maximising the average likelihood $\hat{\theta} = \arg \max_{\theta \in \Theta} \mathcal{L}(\theta|T) / |T|$ where $\mathcal{L}(\theta|T) = \sum_{i=1}^{|T|} \log g(f_{t,C}^i | \theta)$, and $f_{t,C}^i$ is the collection frequency of the i^{th} term $t \in T$.

3.2 Step 4: Best-fitting distribution

In step 4 of the ADR algorithm we select, from the set of candidate statistical distributions, the one that best quantifies the distribution of term frequencies for non-informative terms in C . We do this using Vuong’s Closeness Test [20] and Akaike’s Information Criterion [1], both of which are among the least controversial and widely used statistical tests for model selection [5]. Both tests are based on the Kullback–Leibler divergence and will favour the same model in the limit of large sample sizes, i.e. the model that minimises the information loss w.r.t. the unknown but “true” model.

3.3 Step 5: Distribution-adapted ranking

Let $\hat{\mathcal{M}}$ be the best-fitting distribution to the collection term frequencies of the non-informative terms. Then, in step 5, we replace the distribution P_1 in Eqn. 1 by $\hat{\mathcal{M}}$, yielding:

$$R(q, d) = \sum_{t \in q} f_{t,q} \cdot (-\log_2 \hat{\mathcal{M}}) \cdot (1 - P_2) \quad (3)$$

$\hat{\mathcal{M}}$ captures the same assumption as all DFR models: that non-informative terms are distributed in a certain way. However, whereas in DFR this assumption is not tested, ADR empirically validates the choice of $\hat{\mathcal{M}}$ per collection.

4. EVALUATION

We next evaluate the retrieval performance of our ADR models. As ADR can produce different ranking models for different datasets, we split this section into two parts. Section 4.1 presents our datasets and the ADR models instantiated for these datasets. Section 4.2 compares the retrieval performance of these ADR models against relevant baselines.

4.1 Deriving ADR Models

We use TREC disks 4 and 5 (TREC-d45) with queries 301-450 and 601-700 and ClueWeb cat. B. (CWEB09) with queries 1 - 200, minus query 20 for which no documents are judged relevant as datasets. All datasets are indexed without stop word removal and without stemming using Indri 5.10. For our ADR models, we must identify non-informative terms in each collection and the best fitting distribution of $f_{t,C}$ of the non-informative terms in the collection. We detect non-informative terms using SVM classification, as follows. Given an initial list of 40 informative and non-informative terms (manually compiled by three humans with complete agreement, see Table 1), we compute the term weight of these terms (using IDF [11], x^T [3], residual-IDF (RIDF) [4], gain [15] and z -measure [8] - see [12] for an overview of term weighting), and use these term weights as features to train a binary SVM classifier using 10-fold cross-validation on all feature combinations. We use an SVM approach instead of any of the above term-weighting approaches, as the latter ones require (manual) setting some ad hoc threshold above/below which terms are considered informative/non-informative. In contrast, the best classifier correctly classifies $\approx 86\%$ of the terms using RIDF and gain term weights as features. We use this classifier to classify

each term in TREC-d45 and CWEB09 as informative or non-informative.

Informative terms		Non-informative terms	
hypothermia	intercepted	welcome	awesome
congoese	furloughs	beginner	jolly
outpaced	randomization	spade	out
anthropocentric	existentialist	delete	temp
iridescence	canvass	quit	feels
archdiocesan	colonisation	least	loss
nonconformist	airbrushed	silly	clear
overclocking	leviathans	cent	off
nominalization	inflammation	jar	test
translucent	handmade	chat	fork
shortwave	monasticism	yards	move
crystallography	aperitif	fast	pair
expressionist	pathologize	view	stop
cephalopod	abolishment	fold	colour
paparazzi	bookcase	roll	sit
presided	hydraulic	follow	back
beneficiary	vested	pen	hello
convection	floods	day	stair
custodian	chivalry	flash	best
populist	constrained	money	considerable

Table 1: Informative and non-informative terms.

Next, using steps 3 – 4 of the ADR algorithm we find that the best-fitting statistical model to the distribution of non-informative term frequencies in both datasets is the discrete Yule–Simon (YS) with parameter $p=1.804$ (TREC-d45) and $p=1.627$ (CWEB09).

The YS distribution is defined for $x \in \mathbb{Z}^+, x \geq 1$ [19]:

$$P(x|p+1) = (p+1) \cdot \frac{\Gamma(x) \cdot \Gamma(p+1)}{\Gamma(x+p+1)} \quad (4)$$

where $p > 0$ is the distribution parameter and Γ is the gamma function. Replacing x with $\hat{f}_{t,d}$ (see Section 2), we obtain:

$$P_1^{YS}(\hat{f}_{t,d}|p+1) = (p+1) \cdot \frac{\Gamma(\hat{f}_{t,d}) \cdot \Gamma(p+1)}{\Gamma(\hat{f}_{t,d}+p+1)} \quad (5)$$

which, for a term t , returns the probability of having $\hat{f}_{t,d}$ occurrences of t . Identically to the original Poisson model [2], Eqn. 5 is theoretically dubious as it is a discrete distribution whose input is a real number. Consequently, we use Lanczo’s method to approximate the Γ function identical to [2]. Plugging Eqn. 5 into Eqn. 1 gives:

$$R(q, d) = \sum_{t \in q} f_{t,q} \cdot (-\log_2 P_1^{YS}) \cdot (1 - P_2) \quad (6)$$

We refer to Eqn. 6 as our YS ADR model, or YS for short.

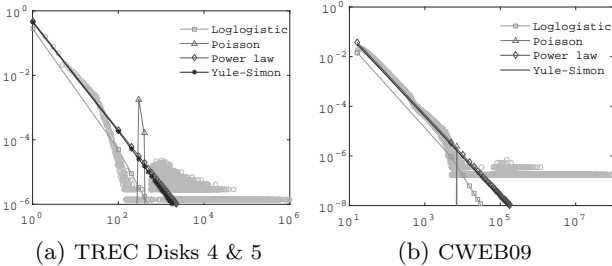


Figure 1: Collection term frequencies for non-informative terms (grey) in CWEB09, and TREC-d45. Superimposed on each figure are MLE fitted log-logistic, Poisson, power law and Yule–Simon distributions as references.

4.2 Retrieval Experiments

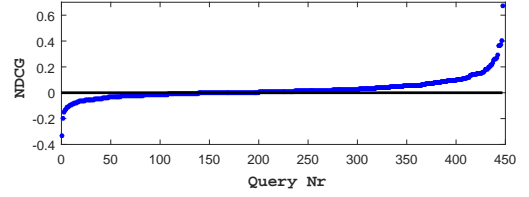


Figure 2: Per query difference in nDCG between YSL2 and LMDir for all queries. Horizontal line indicates no difference. Points above zero favour YSL2 over LMDir.

Baselines and Tuning. We compare our ADR ranking model (YS) to a Poisson (P), tf-idf (I_n), log-logistic (LL), smoothed power law (SPL) and query-likelihood unigram LM with Dirichlet smoothing (LMDir). P and I_n are original DFR models [2]; LL [6] and SPL [7] are information models that extend the original DFR models. All DFR/ADR models are postfixed with L2 meaning that Laplace and logarithmic term normalisation (see Section 2) are used. Following previous work [2, 7], the parameter values of the distributions used in DFR/ADR (e.g. Poisson or Yule–Simon) are induced from the collection statistics in two ways (we experiment with both): $T_{tc} = f_{t,C}/|C|$, or $T_{dc} = n_t/|C|$. E.g., PL2- T_{tc} is a Poisson model using Laplace and logarithmic term normalisation, where Poisson’s $\lambda = T_{tc} = f_{t,C}/|C|$ (see Eqn. 2), and YSL2- T_{dc} is a Yule–Simon ADR model using Laplace and logarithmic term normalisation, where Yule–Simon’s $p = T_{dc} = n_t/|C|$ (see Eqn. 4). We retrieve the top-1000 documents according to each ranking model using title-only queries, and measure performance using P@10, Bpref, ERR@20, nDCG@10 and nDCG. All models are tuned using 3-fold cross-validation and we report the average over all three test folds. We vary c (the free parameter in the logarithmic term normalisation – see Section 2) in the range $\{0.5, 1, 2, 4, 6, 8\}$ [7], and μ of LMDir in the range $\{100, 500, 800, 1000, 2000, 3000, 4000, 5000, 8000, 10000\}$.

Findings. The results are shown in Table 2. For CWEB09, our YSL2 models obtain the best performance at all times. For TREC-d45, our YSL2 models obtain the best performance on early precision retrieval (P@10, ERR@20 and nDCG@10), while the SPL2 models perform best on nDCG and Bpref (with YSL2 following closely). As Fig. 1 shows, the power law (used in SPL2) and Yule–Simon (used in YSL2) have very similar fits, which explains their similar scores in Table 2. In Fig. 1 we also see that the log-logistic, power law and Yule–Simon all approximate the head of the distribution (collection term frequencies up to ≈ 100), though all fail to model the “noisy” tail (collection term frequencies > 1000) well. In contrast, the Poisson distribution fails to model the empirical data in any range, possibly explaining the overall low retrieval performance in Table 2. That the Yule–Simon distribution gives the highest retrieval performance indicates the merit of using the best-fitting distribution for DFR ranking models. Fig. 2 shows the per-query difference in nDCG between the YSL2 and LMDir baseline for all queries. We see that the number of queries that benefit from YSL2 is substantially higher (270 queries) than for the LMDir (163 queries), confirming the scores in Table 2.

5. RELATED WORK

Our work can be seen as a refinement of DFR ranking

	TREC disks 4& 5					ClueWeb09 cat. B.				
Model	nDCG	P@10	Bpref	ERR@20	nDCG@10	nDCG	P@10	Bpref	ERR@20	nDCG@10
LMDir	.4643	.3845	.2239	.1043	.3968	.2973	.2586	.2209	.0973	.1769
PL2- T_{tc} [2]	.2524*	.1273*	.1009*	.0359*	.1332*	.1448*	.0712*	.1258*	.0211*	.0472*
PL2- T_{dc} [2]	.2487*	.1217*	.0960*	.0347*	.1273*	.1444*	.0709*	.1252*	.0314*	.0471*
I_n L2- T_{tc} [2]	.2917*	.1627*	.1114*	.0478*	.1742*	.1596*	.0782*	.1405*	.0352*	.0511*
I_n L2- T_{dc} [2]	.2818*	.1626*	.1088*	.0481*	.1745*	.1596*	.0783*	.1407*	.0352*	.0512*
LLL2- T_{tc} [6]	.4812	.4049	.2341	.1072	.4142	.3184	.2542	.2349	.0926	.1706
LLL2- T_{dc} [6]	.4810	.3982	.2329	.1069	.4097	.3180	.2542	.2349	.0928	.1707
SPLL2- T_{tc} [7]	.4863	.4144	.2375	.1103	.4276	.3207	.2529	.2357	.0945	.1720
SPLL2- T_{dc} [7]	.4876	.4176	.2387	.1107	.4299	.3224	.2586	.2370	.0958	.1752
YSL2- T_{tc} (ADR)	.4644	.3982	.2280	.1048	.4069	.3197	.2601	.2359	.0951	.1752
YSL2- T_{dc} (ADR)	.4860	.4182	.2381	.1113	.4312	.3240	.2666	.2376	.0985	.1810

Table 2: Retrieval performance. Grey denotes larger than the LMDir baseline. Bold marks the best results. * marks statistically significant difference from the LMDir baseline using a t -test at the .05% level.

models. Several other approaches attempt such refinements. For example, Clinchant and Gaussier [6] formulated heuristic retrieval constraints to assess the validity of DFR ranking models, and found that several DFR ranking models do not conform to these heuristics. On this basis, a simplified DFR model based on a log-logistic (LL) statistical distribution was proposed, which adheres to all heuristic retrieval constraints. Experimental evaluation showed that the LL model improved MAP on several collections. In a follow-up study, Clinchant and Gaussier [7] introduced *information models* which are simplified DFR models that do not rely on so-called after-effect normalisation. They instantiated the LL and a “smoothed power law” (SPL) ranking model, and showed that both models tend to outperform a LM, BM25, but not two original DFR models on smaller TREC and CLEF collections. Closest to ours is the work by Hui et al. [10] who developed DFR extensions where the distributions of $f_{t,C}$ are modelled using multiple statistical distributions. However, (i) no justification or empirical validation of the choice of statistical distributions used was given, and (ii) the distribution of all terms, rather than only the non-informative terms, was used. Finally, Hui et al’s models also removed e.g. collection-wide term frequencies from the ranking component, hence effectively removing the notion of divergence from the DFR framework.

6. CONCLUSION

Divergence From Randomness (DFR) ranking models assume that informative terms in a collection are distributed differently than non-informative terms. Different DFR models are produced depending on the statistical model (e.g. Poisson, geometric) used to quantify the distribution of non-informative terms, where an informative term is detected by measuring the divergence of its distribution from the distribution of non-informative terms. However, there is little evidence that the distributions used in DFR actually fit the the distribution of non-informative terms. To address this, we presented a novel DFR extension called Adaptive Distributional Ranking (ADR), which adapts the ranking computation to the statistics of the specific collection being processed each time. Our ADR models first determine the best-fitting distribution of non-informative terms, and then integrate this distribution into ranking. Experiments on TREC data showed that our ADR models outperformed DFR models (and their extensions), and achieved performance comparable to a query likelihood LM. In future work we plan to experiment with additional collections. We will

also study automatically deriving ad hoc distributions suited for the collection data in ADR instead of selecting among a list of standard distributions.

7. ACKNOWLEDGMENTS

Work partially funded by C. Lioma’s FREJA research excellence fellowship (grant no. 790095).

8. REFERENCES

- [1] H. Akaike. A new look at the statistical model identification. *IEEE TAC*, 19(6):716–723, 1974.
- [2] G. Amati and C. J. Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM TOIS*, 20(4):357–389, 2002.
- [3] A. Bookstein and D. R. Swanson. Probabilistic models for automatic indexing. *JASIS*, 25(5):312–316, 1974.
- [4] K. Church and W. Gale. Inverse document frequency (idf): A measure of deviations from Poisson. In *NLP-VLC*, pages 283–295. Springer, 1999.
- [5] K. A. Clarke. Nonparametric model discrimination in international relations. *JCR*, 47(1):72–93, 2003.
- [6] S. Clinchant and É. Gaussier. Bridging language modeling and divergence from randomness models: A log-logistic model for IR. In *ICTIR*, pages 54–65, 2009.
- [7] S. Clinchant and E. Gaussier. Information-based models for Ad Hoc IR. In *SIGIR*, pages 234–241. ACM, 2010.
- [8] S. P. Harter. A probabilistic approach to automatic keyword indexing. part I. on the distribution of specialty words in a technical literature. *JASIS*, 26(4):197–206, 1975.
- [9] S. P. Harter. A probabilistic approach to automatic keyword indexing. part II. an algorithm for probabilistic indexing. *JASIS*, 26(5):280–289, 1975.
- [10] K. Hui, B. He, T. Luo, and B. Wang. Relevance weighting using within-document term statistics. In *CIKM*, pages 99–104. ACM, 2011.
- [11] K. S. Jones. A statistical interpretation of term specificity and its application in retrieval. *JD*, 28(1):11–21, 1972.
- [12] C. Lioma and R. Blanco. Part of speech based term weighting for information retrieval. In M. Boughanem, C. Berrut, J. Mothe, and C. Soulé-Dupuy, editors,

ECIR, volume 5478 of *LNCIS*, pages 412–423. Springer, 2009.

- [13] I. J. Myung. Tutorial on maximum likelihood estimation. *JMP*, 47(1):90–100, 2003.
- [14] I. Ounis, C. Lioma, C. Macdonald, and V. Plachouras. Research directions in terrier: a search engine for advanced retrieval on the web. *Novatica*, VIII(1):49–56, 2007.
- [15] K. Papineni. Why inverse document frequency? In *ACL*, pages 1–8. ACL, 2001.
- [16] C. Petersen, J. G. Simonsen, and C. Lioma. Power law distributions in information retrieval. *ACM TOIS*, 21(4):1–36, 2015.
- [17] J. M. Ponte and W. B. Croft. A language modeling approach to information retrieval. In *SIGIR*, pages 275–281. ACM, 1998.
- [18] S. E. Robertson and S. Walker. Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In *SIGIR*, pages 232–241. Springer, 1994.
- [19] H. A. Simon. On a class of skew distribution functions. *Biometrika*, pages 425–440, 1955.
- [20] Q. H. Vuong. Likelihood ratio tests for model selection & non-nested hypotheses. *Econometrica*, 57:307–333, 1989.