Benchmarking and Improving Recovery of Number of Topics in Latent Dirichlet Allocation Models

Jason Hou-Liu

January 4, 2018

Abstract

Latent Dirichlet Allocation (LDA) is a generative model describing the observed data as being composed of a mixture of underlying unobserved topics, as introduced by Blei et al. (2003). A key hyperparameter of LDA is the number of underlying topics k, which must be estimated empirically in practice. Selecting the appropriate value of k is essentially selecting the correct model to represent the data; an important issue concerning the goodness of fit. We examine in the current work a series of metrics from literature on a quantitative basis by performing benchmarks against a generated dataset with a known value of k and evaluate the ability of each metric to recover the true value, varying over multiple levels of topic resolution in the Dirichlet prior distributions. Finally, we introduce a new metric and heuristic for estimating k and demonstrate improved performance over existing metrics from the literature on several benchmarks.

1 Introduction

Latent Dirichlet Allocation Latent Dirichlet Allocation (Blei et al., 2003) (LDA) is a topic-modeling algorithm suitable for clustering text data. The core principle of LDA is that an observed document is a mixture of some latent underlying topics. Each topic in turn manifests itself as a mixture of some words of the vocabulary of the documents.

Generative Model LDA describes each document to be composed of observable words $w_1, w_2, ..., w_N$, each representing latent unobservable topics $z_1, z_2, ..., z_N$. As a generative hyperparameter, there are k underlying topics throughout the corpus. Here, N is Poisson distributed with some rate λ.

Every document d has a mixture of topics $\vec{\theta}$ drawn from a Dirichlet compound multinomial distribution.

$$\vec{\theta} \sim \text{Dirichlet}(\vec{\alpha} \in \mathbb{R}^k)$$
$$z_i \sim \text{Multinomial}(\vec{\theta})$$

Likewise, every topic j is a mixture of words from the vocabulary $\vec{\phi}_j$ drawn also from a Dirichlet compound multinomial distribution. Let $\mathcal{W}$ represent the size of the vocabulary.

$$\vec{\phi}_{z_i} \sim \text{Dirichlet}(\vec{\beta} \in \mathbb{R}^{\mathcal{W}})$$
$$w_i \mid z_i \sim \text{Multinomial}(\vec{\phi}_{z_i}))$$

Here, $\vec{\alpha}$ and $\vec{\beta}$ are also hyperparameters.

Problem Description One major hurdle to estimating the θ and ϕ parameters is that the number of topics hyperparameter k is assumed to be known a priori. Given a dataset in practice; however, this is typically not the case. Thus, the prevailing course of action is estimate this quantity as a tuning parameter for goodness-of-fit.

There are existing metrics for determining an estimate $\hat{k}$ (Griffiths and Steyvers, 2004; Cao et al., 2009; Deveaud et al., 2014; Arun et al., 2010; Blei et al., 2003) given a series of fitted LDA models. All of these methods require fitting the LDA model multiple times to the same dataset along a series of candidate values of $\hat{k}$. Typically, though it is not required, the series of points are regularly spaced.

In the original LDA paper by Blei et al. (2003), the method of calculating perplexity (inverse of geometric mean of $\log P(w_i)$'s) of a held-out test set is suggested.

Griffiths and Steyvers (2004) require the use of a Gibbs sampler to estimate LDA parameters, and samples the posterior of the Gibbs sampler state at regular intervals and chooses the $\hat{k}$ that maximizes the harmonic mean of the sampled log-likelihoods.

Cao et al. (2009) estimate the average cosine similarity between topic distributions $\vec{\phi}_i$, $\vec{\phi}_j$ $(i \neq j)$ and chooses the value of $\hat{k}$ that minimizes this quantity.

Arun et al. (2010) minimize the symmetric Kullback-Liebler divergence between the singular values of the matrix representing word probabilities for each topic and the topic distribution within the corpus. However, Arun et al. (2010) claims that there is an order

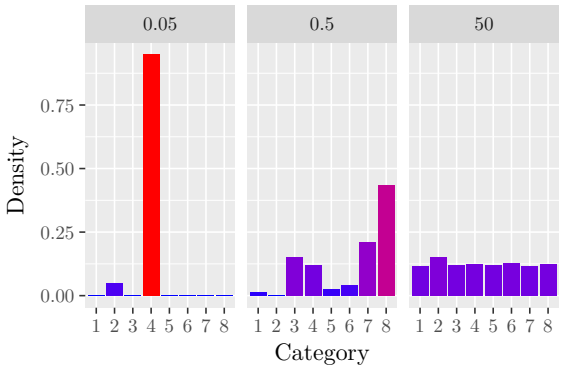

Figure 1: Representative draws from symmetric Dirichlet distributions with $\alpha = 0.05, 0.5, 50$ and eight categories ($n = 8$)

on the singular values of the word-topic matrix, which we find perplexing as singular value decomposition does not produce an implied order of the singular values.

Deveaud et al. (2014) maximize the average Jensen-Shannon distance between all pairs of topic distributions $\vec{\phi}_i$, $\vec{\phi}_j$ $(i \neq j)$, much like Cao et al. (2009).

While there are several methods of determining the optimal number of topics empirically in the literature, a rigorous treatment of their effectiveness is lacking. The present work aims to provide a quantitative measurement of their effectiveness, and introduces an alternative metric with superior recovery properties.

Topic Separability The n-dimensional parameter vector $\vec{\alpha}$ of a Dirichlet distribution is known as a concentration parameter, and can be understood as the weights on the vertices of a standard $(n-1)$ simplex in $\mathbb{R}^{n-1}$. In the particular case of the symmetric Dirichlet distribution, all n elements of the parameter vector are equal, and thus the symmetric Dirichlet distribution is fully characterized by a single scalar parameter α.

The key interpretation of the parameter that decreasing α causes the weight of the density function becomes concentrated at the vertices of the simplex. When α increases, the density function becomes concentrated in the interior (Figure 1).

As a prior for the topic and word multinomial distributions in LDA, a larger parameter value for the Dirichlet prior produces poorly resolved topics; documents become increasingly ambiguous about their topic mixture, and topics increasingly share words with other topics.

2 Methods

Datasets The datasets used in the current work are created as per the description of the generative model. As the application of Gibbs sampling in LDA requires symmetric prior hyperparameters (Griffiths and Steyvers, 2004), we take the values α and β to be 0.05, 0.1, and 0.2, representing well-separated, separated, and poorly separated topics, respectively. In all three cases, both α and β are chosen to be equal.

To evaluate recovery of the number of topics at different levels of topic separability, datasets for different values of k are generated for each set of hyperparameters. In particular, the following values for k are chosen:

$$k = 10, 15, ..., 95, 100$$

Each dataset consists of 1000 documents sharing a $\mathcal{W} = 1000$ word vocabulary. In each case, the chosen Poisson rate parameter λ is 100. To mitigate spurious effects of a single run, ten datasets are generated under each set of conditions.

The following pseudo-code describes the generative process algorithmically:

for $d = 1$ **to** 1000 **do**
  sample $N \sim \text{Poisson}(\lambda)$
  for $i = 1$ **to** N **do**
    sample $z_i^{(d)} \sim \text{Multinomial}(\vec{\theta}_{(d)})$
    sample $w_i^{(d)} \sim \text{Multinomial}(\vec{\phi}_{(z_i)})$
  end
end

Model Fitting While the original paper by Blei et al. (2003) describes the process of using Variational Expectation-Maximization to accomplish inference of the model parameters, we use instead the process of Gibbs sampling as described in Griffiths and Steyvers (2004) throughout the current work. In particular, this allows the use of the metric defined in Griffiths and Steyvers (2004). Computationally, we use the `topicmodels` package from `CRAN` (Grün and Hornik, 2011), which relies on the `GibbsLDA++` library (Phan et al., 2008; Phan and Nguyen, 2007).

For each dataset, candidate values $\hat{k}$ are chosen to be:

$$\hat{k} = 5, 10, 15, ..., 195, 200$$

The Gibbs sampling method of fitting LDA models is applied to each combination of (α, β), k, and $\hat{k}$. An upper limit of 2000 iterations is used, and in the case of Griffiths and Steyvers (2004), every 50^{th} iteration is sampled to prevent sampling from correlated states of the Markov chain.

To simulate the Dirichlet hyperparameters α and β being unknown, we use $\hat{\alpha} = 50/\hat{k}$ and $\hat{\beta} = 0.1$ when fitting the LDA as suggested in Griffiths and Steyvers (2004).

Evaluation For each dataset, the estimated number of topics k^* is selected according to the criteria of each metric. Values for the existing literature metrics are calculated using code adapted from the `ldatuning` package on `CRAN` (Nikita, 2016). With each set of hyperparameters, we plot the estimated k^* from each metric in each run against the true value of k corresponding to that dataset.

3 New Metric

We propose in the current work a new method to estimate the value $\hat{k}$. We use the same notation as found in Blei et al. (2003):

$$\theta_j^{(d)} = P(z = j \mid d) \text{ in document } d$$
$$\phi_w^{(j)} = P(w \mid z = j)$$

Subsequently, we estimate the probability of an observed document $P(d)$ in the corpus (Heinrich, 2005):

$$\pi_d \propto \left[\prod_{w_i} \left(\sum_j \theta_j^{(d)} \phi_{w_i}^{(j)} \right) \right]^{\frac{1}{\|w_i\|}}$$

Note that $\propto$ is used since there is no guarantee that the sum of π_d over all observed documents will be unity. Normalization is induced by dividing each π_d by $\sum_d \pi_d$.

Using conditional probability:

$$P(z = j) = \sum_d \pi_d \theta_j^{(d)}$$
$$P(w) = \sum_{j,d} \pi_d \phi_w^{(j)} \theta_j^{(d)}$$
$$P(z = j \mid w) = \phi_w^{(j)} \frac{\sum_d \pi_d \theta_j^{(d)}}{\sum_{j,d} \pi_d \phi_w^{(j)} \theta_j^{(d)}}$$

And again:

$$P(d \mid z = j) = \frac{\pi_d \theta_j^{(d)}}{\sum_d \pi_d \theta_j^{(d)}}$$

Estimating the topic distribution for document d with words $w_1, w_2, ..., w_N$ using the above probabilities:

$$\hat{P}(z = j \mid d) \propto \prod_{w_i} P(z = j \mid w_i)$$
$$= \prod_{w_i} \left(\phi_{w_i}^{(j)} \frac{\sum_d \pi_d \theta_j^{(d)}}{\sum_{j,d} \pi_d \phi_{w_i}^{(j)} \theta_j^{(d)}} \right)$$

Again, $\propto$ is used since there is no guarantee that the sum over topics $j = 1, 2, ..., \hat{k}$ will be unity. Normalization is again enforced by dividing through by $\sum_j \hat{P}(z = j \mid d)$. Subsequently, for any document d' in the corpus, calculate the confusion between document d and d' as:

$$\hat{P}(d' \mid d) = \sum_j \left[P(d' \mid z = j) \cdot \hat{P}(z = j \mid d) \right]$$
$$= \sum_j \left[\frac{\pi_{d'} \theta_j^{(d')}}{\sum_d \pi_d \theta_j^{(d)}} \prod_{w_i} \left(\phi_{w_i}^{(j)} \frac{\sum_d \pi_d \theta_j^{(d)}}{\sum_{j,d} \pi_d \phi_{w_i}^{(j)} \theta_j^{(d)}} \right) \right]$$

Define the confusion matrix $\mathbf{C}|\hat{k}$ under the candidate number of topics $\hat{k}$ element-wise as:

$$\{c_{ij}|\hat{k}\} = \hat{P}(d_i|d_j)$$

Each column $\vec{c_j}|\hat{k}$ of $\mathbf{C}|\hat{k}$ represents the confusion of document d_j with all other documents of the corpus. Consider that in the ideal case, there would be no confusion, and so $\vec{c_j}|\hat{k} = \vec{e_j}$ where $\vec{e_j}$ is the j^{th} elementary basis vector of $\mathbb{R}^{\mathcal{D}}$. Hence, the ideal confusion matrix is simply the $\mathcal{D} \times \mathcal{D}$ identity matrix:

$$\{c_{ij}|\hat{k}\} = \delta_{ij}$$

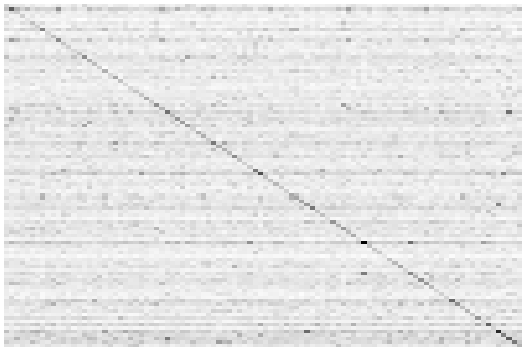

Figure 2: Leading principal minor of order 100 of a confusion matrix. Darker cells indicate higher values.

Figure 2 displays the leading principal minor of order 100 of a confusion matrix as calculated using the steps described above. The main diagonal visibly contains more weight, emulating the identity matrix. Note that while it is intuitive, the confusion matrix $\mathbf{C}|\hat{k}$ only suggests that confusing document d_i with d_j means also

confusing document d_j with d_i; it is not a truly symmetric matrix.

Finally, define the metric's scoring function as the average cosine similarity of the columns $\vec{c_j}|\hat{k}$ with $\vec{e_j}$:

$$M(\hat{k}) = \frac{1}{\mathcal{D}} \sum_{j=1}^{\mathcal{D}} \frac{\vec{c_j} \cdot \vec{e_j}}{\|\vec{c_j}\|_2 \|\vec{e_j}\|_2}$$
$$= \frac{1}{\mathcal{D}} \sum_{j=1}^{\mathcal{D}} \frac{c_{jj}}{\|\vec{c_j}\|_2}$$

Elbow Heuristic While the optimum value k^* may be estimated visually using the elbow method for each dataset, a simple approximate heuristic is provided below for identifying k^* automatically with the reasonable assumption that the candidate values $\hat{k}$ contains the true value k. Ideally, k will be contained reasonably interior in the set of $\hat{k}$'s.

Suppose $\hat{k} = \hat{k}_1, \hat{k}_2, ..., \hat{k}_n$ and that for each $\hat{k}_i$ there is a corresponding metric value $M(\hat{k}_i)$. Fit a local regression (LOESS) using $M(\hat{k}_i)$ against $\hat{k}_i$. A span of 0.75 works reasonably well here. We select the estimated elbow k^* as the maximum positive residual from this regression:

$$k^*_{\text{New2017}} = \max_{1 \le i \le n} \left[\hat{M}(\hat{k}_i) - M(\hat{k}_i) \right]$$

4 Results

Figure 3 displays the estimated number of topics k^* against the true number of topics k. The dark grey line $k^* = k$ is underlaid to indicate a correct recovery of the quantity k. A metric should ideally follow this line exactly at all points; signifying correct estimation of k.

Immediately, it is evident that Arun et al. (2010) and Deveaud et al. (2014) produce poor estimates throughout. While Deveaud et al. (2014) identifies low k correctly in the case of $\alpha = \beta = 0.05$, it grossly undershoots in all other situations. In contrast, Arun et al. (2010) grossly overshoots at all tested values of k. Further analyses will ignore these two metrics as they seem to be uninformative on the majority of the generated datasets.

For a closer look at the behaviors of each metric, Figures 4, 5, and 6 display the actual metric values of three remaining metrics averaged over the ten replicates for each combination of k and $\hat{k}$. The true value of k is marked on each graph on the corresponding line as an hollow circle, and the chosen k^* for each metric is a solid circle. Note that the displayed k^* is not derived after averaging the metric values, but is the average of the k^* of each dataset. A correct recovery of k (on average) would thus be represented as the solid circle being 'on-target'.

Figures 4, 5, and 6 have the effect of isolating the effectiveness of the actual metric by smoothing out deviations caused by the idiosyncrasies of each dataset.

Well-Separated Case In the case of Dirichlet priors $\alpha = \beta = 0.05$ (Figure 4), Griffiths and Steyvers (2004) recovers the true value of exactly for $k = 35, 40, ..., 75$, with appreciable performance for larger values of k. Cao et al. (2009) does similarly; exact recovery for $k = 25, 30, ..., 60$ and similarly appreciable performance for larger k. However, both these metrics display signs of overfitting below their comfort zone, more severely so for Griffiths and Steyvers (2004) than Cao et al. (2009). The new metric proposed in this paper suffers much less overfitting at low values of k, and maintains good performance throughout the tested range of k. All three metrics display initial signs of underfit as k exceeds 80, particularly for Cao et al. (2009).

The new metric displays a clear kink at the true number of topics. Further, for lower number of topics $k = 10, 15, 20$, the curve displays a noticeable bump after the kink with maxima consistent with the overfit pitfalls suffered by Griffiths and Steyvers (2004) and Cao et al. (2009). The conclusion is that the new metric is able to both identify the correct value of k here, as well as provide evidence for potential overfit.

Semi-Separated Case When the separatedness begins to decrease, at $\alpha = \beta = 0.1$ (Figure 5), significant deviations in the ability to recover k appear. The new metric maintains superior recovery at lower values of k, but suffers the same reduction as the other two at higher k.

While the 'sweet spot' for Griffiths and Steyvers (2004) extends higher than the new metric, it is considerably narrower. Cao et al. (2009) shows similar behavior.

Furthermore, the new metric demonstrates more regular spacing of the solid dots for $k > 60$ than the other two metrics, which exhibit irregular jumps. This suggests that the new proposed metric is more stable when the dataset is drawn from a less separated distribution.

Least-Separated Case Finally, in the least separated case considered in this paper (Figure 6), there is a clear breakdown of topic number recoverability across all three metrics.

The workable range for Griffiths and Steyvers (2004) reduces considerably, and Cao et al. (2009) suffers a less dramatic contraction. This paper's metric gains

similar characteristics as Griffiths and Steyvers (2004); the pattern of underfit appears very similar. Both seem to experience a moderate decrease in k^* at $k = 30/35$

Additional Observations From Figure 3, Griffiths and Steyvers (2004) appears to have the least sensitivity to idiosyncracies in individual datasets; at each value of k, the dots remain near the groupwise mean. For Cao et al. (2009) and the new metric of this paper, the variance increases outside of their effective recovery ranges.

An interesting interpretation of the effects seen when $\alpha = \beta = 0.1$ is that adding incremental topics above 60 is not compatible with the nature of the generated dataset. In other words, 1000 documents and a 1000-word vocabulary simply cannot support more than 60 topics at this level of separability. Consequently, all three metrics maybe preferring to merge similar topics, with additional incremental topics serving only to chain-link many topics into one large topic.

Extending this interpretation to the least-separated case of $\alpha = \beta = 0.2$, a similar effect is seen. though k^* seems to stabilize at around 25 to 35 as k approaches 100. It stands to reason that a finite corpus of documents with a finite length and vocabulary will have a limit of sustainable topics; much as the maximum resolvable frequency is half the sampling rate in waveforms, there is an upper bound on the number of topics carried in a corpus.

Additionally, Cao et al. (2009) overshoots relative to Griffiths and Steyvers (2004) in the semi-separated case but the converse becomes true in the least-separated case. This seems to suggest a strong sensitivity to poorly separated topics in the metric of Cao et al. (2009); it remains to be seen how the behavior changes at even higher values of the hyperparameters α and β.

5 Conclusion

We show in the present work that there is room for improving upon existing metrics for identifying the optimum number of topics in the literature of Latent Dirichlet Allocation. This work also demonstrates Arun et al. (2010) and Deveaud et al. (2014) erring significantly in the direction of overfit and underfit, respectively. We demonstrate our new metric to have superior recovery of k when the true number of topics is low, outperforming Griffiths and Steyvers (2004) and Cao et al. (2009) in situations pervious to underfit, while maintaining comparable performance elsewhere.

References

D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of machine Learning research*, vol. 3, no. Jan, pp. 993–1022, 2003.

T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.

J. Cao, T. Xia, J. Li, Y. Zhang, and S. Tang, "A density-based method for adaptive lda model selection," *Neurocomputing*, vol. 72, no. 7, pp. 1775–1781, 2009.

R. Deveaud, E. SanJuan, and P. Bellot, "Accurate and effective latent concept modeling for ad hoc information retrieval," *Document numérique*, vol. 17, no. 1, pp. 61–84, 2014.

R. Arun, V. Suresh, C. V. Madhavan, and M. N. Murthy, "On finding the natural number of topics with latent dirichlet allocation: Some observations," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2010, pp. 391–402.

B. Grün and K. Hornik, "topicmodels: An R package for fitting topic models," *Journal of Statistical Software*, vol. 40, no. 13, pp. 1–30, 2011.

X.-H. Phan, L.-M. Nguyen, and S. Horiguchi, "Learning to classify short and sparse text & web with hidden topics from large-scale data collections," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 91–100.

X.-H. Phan and C.-T. Nguyen, "Gibbslda++: Ac/c++ implementation of latent dirichlet allocation (lda)," *Tech. rep.*, 2007.

M. Nikita, *ldatuning: Tuning of the Latent Dirichlet Allocation Models Parameters*, 2016, r package version 0.2.0. [Online]. Available: https://CRAN.R-project.org/package=ldatuning

G. Heinrich, "Parameter estimation for text analysis (2005)," *Web: http://www.arbylon.net/publications/text-est2.pdf*, 2005.

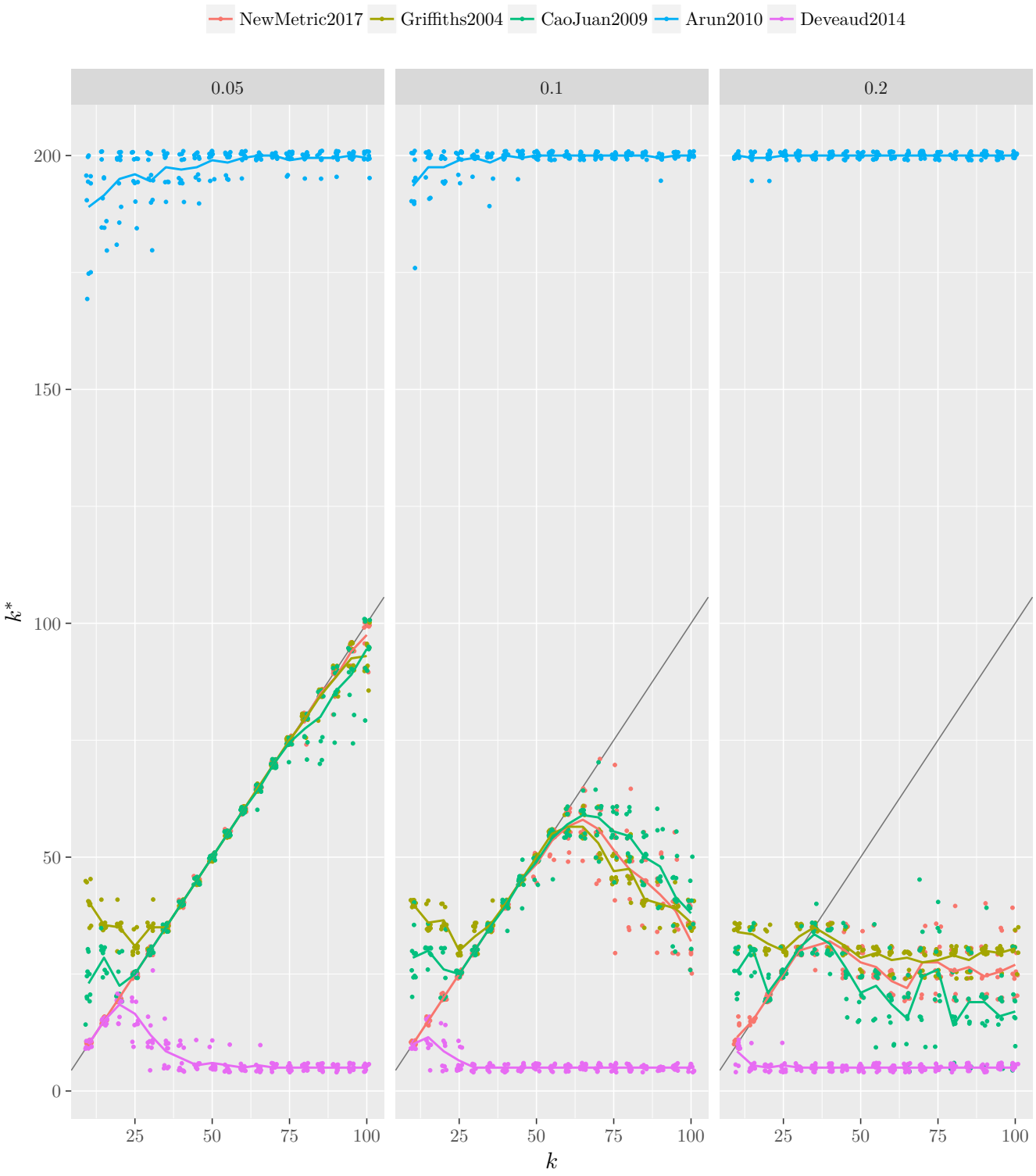

Figure 3: Comparison between fitted values k^* and true values k across all five different tested metrics. The dark grey line represents the 45° line $k^* = k$. Headings for each of the three graphs denote the value of the Dirichlet hyperparameters α and β. Each coloured dot represents the estimated value $k^*_{(\cdot)}$ of a single dataset for metric $(\cdot)$. Metric names are drawn from the labels in the `ldatuning` package (Nikita, 2016).

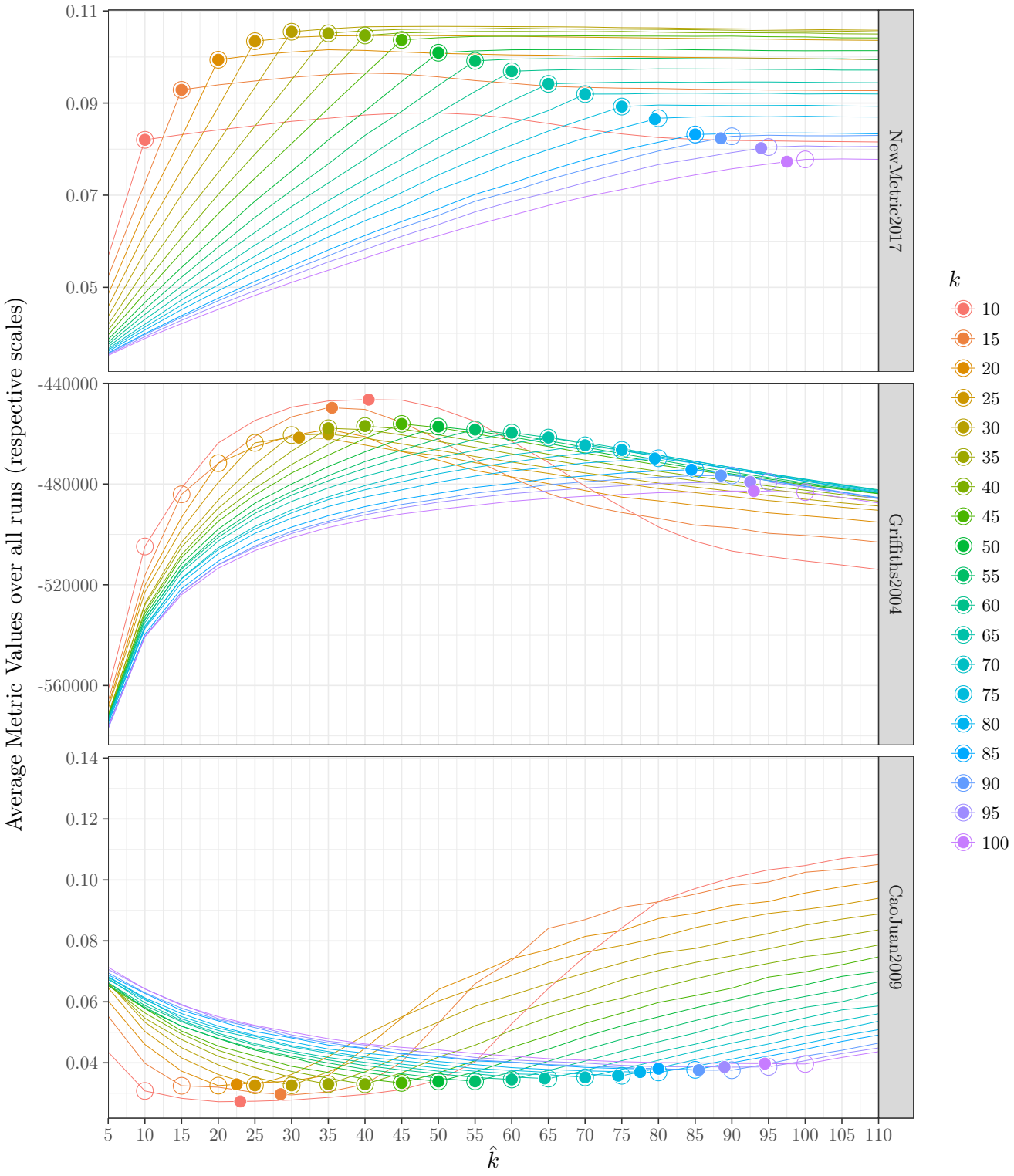

Figure 4: Computed metrics and estimated k^* under Dirichlet hyperparameters $\alpha = \beta = 0.05$. Each line represents an average over 10 runs of each value of k. Lines are interpolated between the averages across all runs at each tested value $\hat{k}$. An open circle denotes the position of the true value of k on the corresponding line. A closed circle denotes the average of the estimated k^* across all runs for the corresponding line.

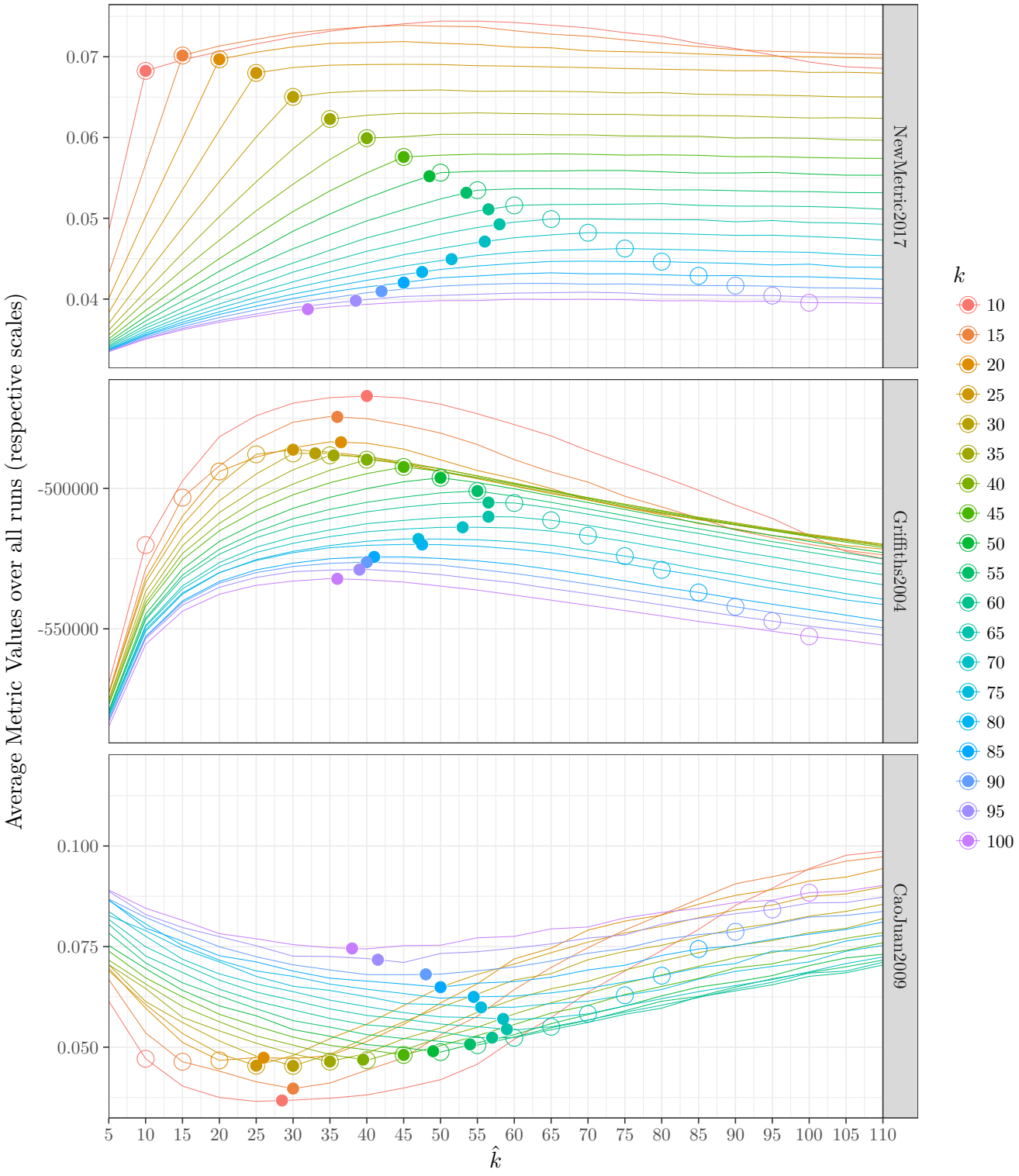

Figure 5: Computed metrics and estimated k^* under Dirichlet hyperparameters $\alpha = \beta = 0.1$. Each line represents an average over 10 runs of each value of k. Lines are interpolated between the averages across all runs at each tested value $\hat{k}$. An open circle denotes the position of the true value of k on the corresponding line. A closed circle denotes the average of the estimated k^* across all runs for the corresponding line.

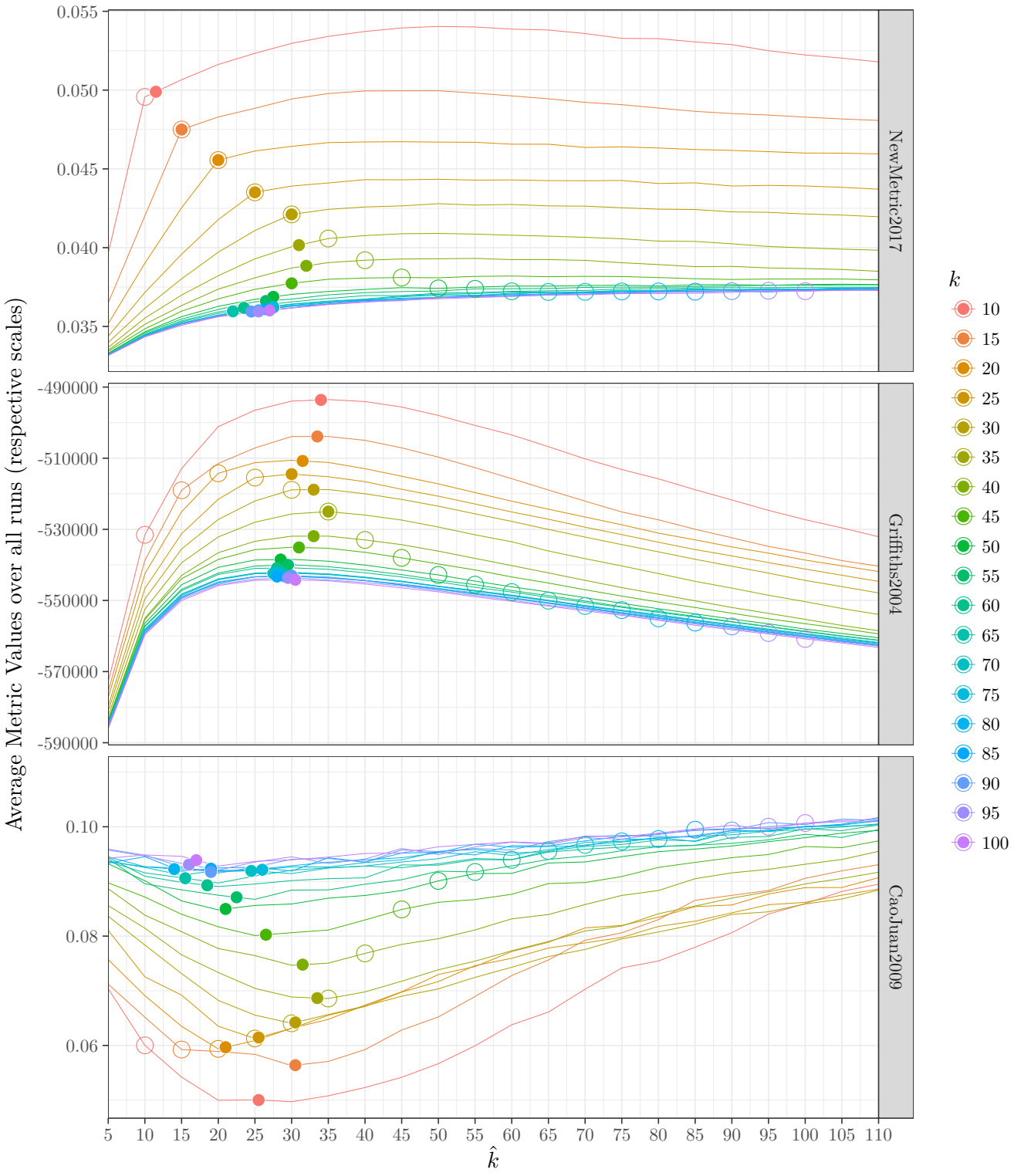

Figure 6: Computed metrics and estimated k^* under Dirichlet hyperparameters $\alpha = \beta = 0.2$. Each line represents an average over 10 runs of each value of k. Lines are interpolated between the averages across all runs at each tested value $\hat{k}$. An open circle denotes the position of the true value of k on the corresponding line. A closed circle denotes the average of the estimated k^* across all runs for the corresponding line.