Margaret R. Donald* and Susan R. Wilson

Comparison and visualisation of agreement for paired lists of rankings

DOI 10.1515/sagmb-2016-0036

Abstract: Output from analysis of a high-throughput 'omics' experiment very often is a ranked list. One commonly encountered example is a ranked list of differentially expressed genes from a gene expression experiment, with a length of many hundreds of genes. There are numerous situations where interest is in the comparison of outputs following, say, two (or more) different experiments, or of different approaches to the analysis that produce different ranked lists. Rather than considering exact agreement between the rankings, following others, we consider two ranked lists to be in agreement if the rankings differ by some fixed distance. Generally only a relatively small subset of the k top-ranked items will be in agreement. So the aim is to find the point k at which the probability of agreement in rankings changes from being greater than 0.5 to being less than 0.5. We use penalized splines and a Bayesian logit model, to give a nonparametric smooth to the sequence of agreements, as well as pointwise credible intervals for the probability of agreement. Our approach produces a point estimate and a credible interval for k. R code is provided. The method is applied to rankings of genes from breast cancer microarray experiments.

1 Introduction

Ranked lists and comparisons between ranked lists have been of interest for a considerable time [see e.g. Mallows (1957)], while sequences representing agreement data have also received attention (Stevens, 1939; Mood, 1940). In the 'omics' literature, ranked lists have become particularly important, given that classical probability theory is not easily applicable. Those not wishing to arbitrarily partition data by using particular cutoffs for p-values or false discovery rates (Eden et al., 2007) rank their data. Having determined the top two hundred or so genes, a common practice is then to import them into network software (for example, the open source software Cytoscape (Shannon et al., 2003) for further analysis. The question arises: Are the top genes meaningfully ranked?

Methods of comparing ranked lists have led to the development of many packages for considering paired ranked lists (Lottaz et al., 2006; Eden et al., 2009; Plaisier et al., 2010; Antosh et al., 2013; Pihur et al., 2014; Schimek et al., 2015b). Hall and Schimek (2012) (referred to in the following as H-S) provide an excellent review of the statistical literature on ordered lists.

H-S propose that lists of rankings degenerate at some point into noise, and find that point (k) for paired lists of rankings. We take an alternative approach and provide (i) credible intervals for the point (k) where sequences of rankings degenerate into noise, and (ii) a visualization of the probability of agreement.

The H-S algorithm is implemented in the R package, TopKLists (Schimek et al., 2015b). We use two key ideas from H-S:

1. Agreements between rankings degenerate at some point into noise;
2. Rank agreement should be more loosely defined than exact agreement in rank.

Thus, we follow H-S in defining agreement between ranks as being based on a distance between ranks, δ, and look for the point k at which the probability of agreement in rankings changes from being greater than 0.5 to being less than 0.5. We approach the problem via penalized splines and generalized linear models in

***Corresponding author: Margaret R. Donald,** Stats Central, University of New South Wales, Anzac Parade, Kensington, NSW, 2052, Australia, e-mail: m.r.donald@bigpond.com
Susan R. Wilson: Stats Central, University of New South Wales, Anzac Parade, Kensington, NSW, 2052, Australia; and Australian National University, Canberra, ACT, 0200, Australia

Table 1: Breast cancer data from TopKLists (Schimek et al., 2015a).

Ranking label	PubMed ID	GEO ID	References
TransBIG	17545524	GSE7390	Desmedt et al. (2007)
MDCC	20676074	GSE20194	MAQC Consortium (2006)
Pusztai	20829329	GSE20271	Tabchy et al. (2010)

an MCMC framework, since this gives a mechanism for finding a non-parametric smooth to the sequence of agreements, while giving pointwise credible intervals for the probability of agreement at any point.

2 Methods

2.1 Data

To illustrate the method, we use the rankings of genes from three breast cancer microarray experiments supplied as a dataset in the R package, TopKLists (Schimek et al., 2015b), and used in the corresponding vignette (Schimek et al., 2015a). Rankings are those given in the dataset, and are not those of the original sources. Table 1 shows the data sources.

2.2 Agreement in rankings

Two ranked lists are considered to be in agreement if the rankings differ by some fixed distance. Thus, in two ranked lists of say 10,000 objects, a difference in rank, of say, up to and less than 100, may be considered agreement in rank. For example, an item ranked 2003rd in one list might be considered to agree in ranking with an item in another list where it ranks 2103. In a list of 200 objects, a reasonable difference in rank might be smaller, e.g. a difference of five, where an item ranked fifth in one list would be considered to agree with the ranking in the second list where it ranks tenth. This difference in ranking agreement appears in the H-S algorithm (and in TopKLists) as the distance parameter, δ.

Consider two ranked lists, $L1$ and $L2$. Suppose $L1$ consists of the objects A, B, C, D, E, F, G in that order, while $L2$ consists of D, B, C, E, A, G, F in that order. Then, taking the ordering of the sequence of agreements from $L1$, and using the exact agreement ($\delta = 0$) of the ranks, the sequence of agreements is 0, 1, 1, 0, 0, 0, 0 where one represents agreement, and zero, disagreement. If the rankings are considered to be in agreement when they deviate by a deviation of $\delta = 1$ in rank, the sequence of agreements becomes 0, 1, 1, 0, 1, 1, 1. The process of forming a sequence of agreements is asymmetric, with the sequence of agreements being taken from the first list.

More generally, let us suppose if an item is ranked j on the first list, then $I_j = 1$, if on the second list the item is not more than δ ranks distant from j, otherwise $I_j = 0$. Thus, $I_1, I_2, \ldots, I_N$, represents the sequence of agreements between the rankings from List 1 ($L1$) and List 2 ($L2$), with the order of the I_j being based on $L1$. Let p_j be the probability of agreement between rankings at the jth sequence point. The H-S algorithm finds the value of k, the point just before the sequence degenerates into noise (defined as the point where p_j becomes less than 0.5). H-S also assume that the decrease of p_j for increasing j is not necessarily monotone.

2.3 Penalized spline fit

We use a generalized additive model (Hastie and Tibshirani, 1990) and fit the data as a logistic curve via generalized linear modelling (McCullagh and Nelder, 1989; Dobson and Barnett, 2008). These techniques are well explored, with many R-packages supporting them, e.g. “nlme” (Pinheiro et al., 2016). In the context

of such modelling, Wand and Ormerod (2008) demonstrated the usefulness of O'Sullivan penalized splines (O'Sullivan, 1986) for semiparametric regression, advocating their use since they are (a) similar to the widely used P-splines, (b) a direct generalization of smoothing splines, (c) have natural boundary properties, and (d) are computationally robust. Three further papers, (Crainiceanu et al., 2005; Wand, 2009; Marley and Wand, 2010) give code for fitting such models in a Bayesian context using the R package BRugs (Thomas et al., 2006).

We follow Wand and Ormerod (2008), and use a Gibbs sampler to fit the splines. The sequence, $j = 1, 2, \ldots, N$, is the predictor variable and is standardized to give a mean of zero and a standard deviation of one. Centering allows better mixing of the MCMC chains, because it reduces collinearity of the intercept with the other coefficients in the model, and standardizing to a standard deviation of one means that priors for coefficients do not have to be tailored to particular datasets (Lunn et al., 2013). This gives the vector $\mathbf{x}$ of length N.

We use a Bayesian logit model. Let $\mathbf{y}$ be the vector of agreements of length N, $\mathbf{X}$ be a matrix, consisting of the concatenation of a vector of ones, and the vector $\mathbf{x}$, the vector of the sequence $1:N$, which has been standardized, giving $\mathbf{X} = [1_{N\times 1}, x_{N\times 1}]$, an $N \times 2$ matrix. Then the model is:

$$\begin{cases} \mathbf{y} \sim \text{Bernoulli}(\mathbf{p}), \\ \text{logit}(\mathbf{p}) = \tilde{\mu}, \\ \tilde{\mu} = \mathbf{X}\tilde{\beta} + \mathbf{Zu}, \\ \tilde{\beta} \sim N(0, \sigma_b^2), \\ \mathbf{u} \sim N(0, \sigma_u^2), \\ \sigma_u \sim \text{Half-Cauchy}(A) \\ \sigma_b^2 = 10{,}000 \end{cases}$$

where $\mathbf{Z}$ is the $N \times (\kappa + 2)$ matrix of the O'Sullivan penalized spline bases for κ internal knots. The first three terms of the model describe a standard generalized linear model (McCullagh and Nelder, 1989). The Gibbs sampler estimates the coefficients β, and $\mathbf{u}$, together with a common variance for the penalty spline coefficients, σ_u^2, which sets the penalty. The spline penalty, λ, is $1/\sigma_u$. The final terms in the model are unconditional, uninformative, priors for σ_u^2 and σ_b^2. Parameters, including λ (and credible intervals) are found from their posterior distributions. (Note, in the Gibbs sampler, samples for each term are drawn conditional on all other terms in the model and the data.)

The smooth function, $p_j = 1/(1 + \exp(-\mu_j))$ where $\tilde{\mu} = \mathbf{X}\tilde{\beta} + \mathbf{Zu}$ forms the basis for finding k and its credible intervals. Technical details for the calculation of k and the effective degrees of freedom (*edf*) of the spline smooth are found in Appendix A.

We provide an R code vignette (*RankAgreeVignetteV10.pdf*) together with a script (*VignetteScript4.R*) in online supplementary materials.

3 Results: comparisons for TopKLists breast cancer data

Table 2 shows estimates for k and its 95% credible intervals using the penalized spline method for the agreement of the comparison between the rankings for TransBig vs MDCC with the full sequence length ($N = 917$), and also the estimates for k using the TopKLists method. For each agreement distance, δ, the values from using the penalized spline method are effectively constant despite the varying degrees of freedom (df) (13–117). This indicates that estimates of k are not sensitive to the fitted number of degrees of freedom. Figure 1 shows the fit when $\delta = 6$, df $= 21$. Figures 2 and 3 show fits when $\delta = 100$, for $N = 917$ and 100 respectively.

The penalty, λ, is a global parameter, and thus, might be thought to have an undue influence on behaviour of the curve at the beginning of the sequence, where interest centres. Hence, we looked at the effect of fitting the penalized splines over shorter agreement sequences (where agreement has previously been calculated across the entire sequence of paired rankings). Table 3 shows the effect of fitting different sequence lengths to the same sequence agreement of TransBig vs MDCC. (Some of these fits push the limits of the penalized spline: fitting splines with 19 degrees of freedom to 50 points allows fewer than three data points between

Table 2: Summary of fits to the breast cancer data: TransBig vs. MDCC.

δ	Method	df	v	k	95% CI
6	Penalized spline	13		0	(0, 18)
		17		0	(0, 18)
		21		0	(0, 18)
		27		0	(0, 18)
		51		0	(0, 18)
	TopKLists		10	10	
10	Penalized spline	13		0	(0, 23)
		17		0	(0, 22)
		21		0	(0, 22)
		27		0	(0, 22)
		51		0	(0, 22)
	TopKLists		10	10	
40	Penalized spline	13		37	(4, 55)
		17		37	(4, 56)
		21		37	(4, 56)
		27		37	(4, 56)
		51		37	(3, 56)
	TopKLists		20	17	
			40	23	
100	Penalized spline	13		97	(73, 123)
		17		97	(73, 124)
		21		97	(73, 124)
		27		96	(72, 124)
		51		96	(72, 123)
		117		97	(72, 123)
	TopKLists		20	34	
			30	73	
			40	78	
			70	67	
			100	73	

Fits are for the full sequence length, $N = 917$. See text for explanations of notation.

knots.) For these data, the credible intervals generally vary by relatively small amounts for the chosen δ. However, point estimates for k increase markedly as N decreases (for $6 \leq \delta \leq 20$) but not for $30 \leq \delta \leq 100$. The lower bounds of the credible intervals change for $40 \leq \delta \leq 100$. The upper limit when $N = 100$ and $\delta = 100$ should be ignored since the upper bound is constrained by the length of the fitted sequence, and is therefore meaningless. (When sequences are truncated, estimates for k and its credible intervals are generally not affected by the truncation since they are estimated as the quantiles of the posterior distribution.) For this sequence, we note that for $\delta = 100$ TopKLists using $v = 20$ gives k as 34, but with $v = 40$ gives k as 78 (Table 2), where v is the window as used by H-S for finding estimates for p_j.

Tables 4 and 5 show the effects of fitting different sequence lengths (and different numbers of degrees of freedom) to the remaining pairs of sequences in the breast data of TopKLists (MDCC vs. Pusztai, and TransBig vs. Pusztai, respectively). Table 4 (MDCC vs. Pusztai) shows point estimates which (1) increase with decreasing fitted sequence length (N) for $\delta = 6, 10, 20$; (2) do not increase for $\delta = 30$; and (3) decrease when $\delta = 40$ and 100. Upper bounds of the 95% CIs may increase ($\delta = 20$) or decrease ($\delta = 100$) with decreasing N, but are generally consistent. Similarly, lower bounds may increase ($\delta = 6, 10, 20$) or decrease ($\delta = 40, 100$). Figure 4 ($\delta = 100$, $N = 150$, df = 19) shows additional bumps compared with Figure 5 ($\delta = 100$, $N = 917$, df = 15), which has a considerably smoother curve, illustrating the underlying reasons for the observed differences. Table 5 shows similar patterns of differences: point estimates increase for decreasing fitted sequence length (N), but credible intervals change very little for $\delta = 6$, and 10; for $\delta = 40$, lower bounds and point estimates

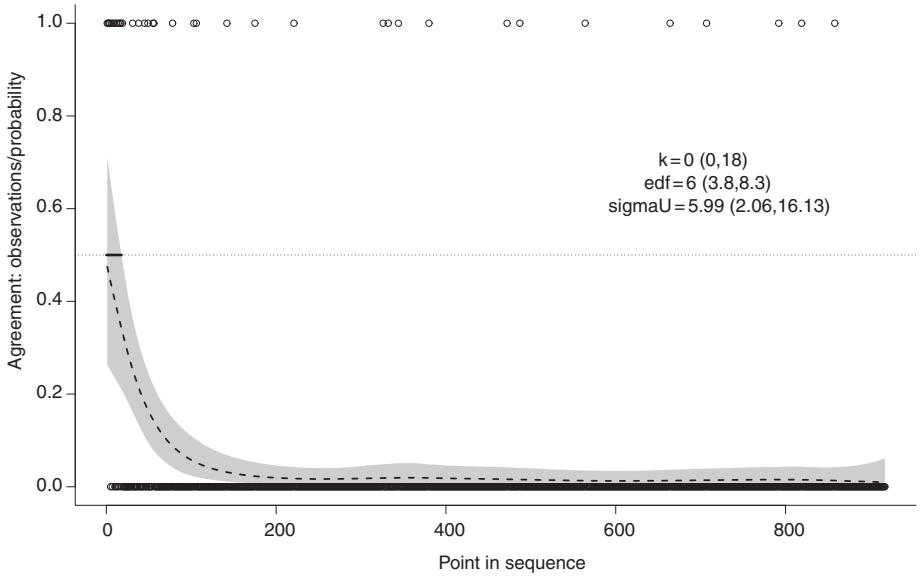

Figure 1: Observed agreement (1s and 0s) and probability of agreement for the TransBig and MDCC rankings, from the breast cancer data from TopKLists ($N = 917$, $\delta = 6$, 21 df.) The dashed curve shows the posterior median for the probability of agreement. The (short) solid line shows the 95% credible interval for k. See the text and Apppendix A for technical details.

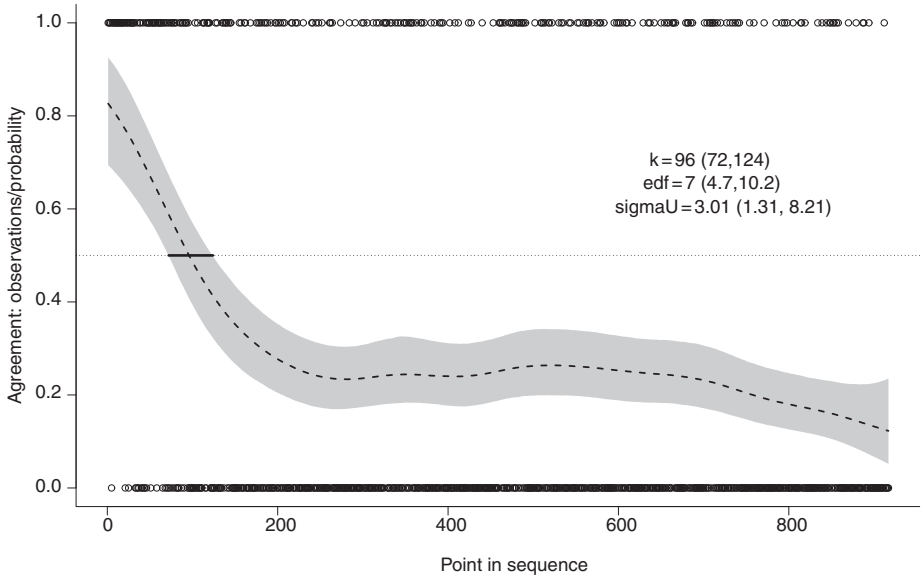

Figure 2: Same as Figure 1 with $\delta = 100$, 117 df. The dashed curve shows the posterior median for the probability of agreement.

increase with decreasing N; while for $\delta = 100$, point estimates decrease and credible intervals become tighter with decreasing N. The patterns observed are dictated by the agreement sequence, with the extra lumps and bumps found in the shorter sequences (as seen in Figures 3 and 4) changing the way in which the probability curve crosses the probability value of 0.5. However, it is always the case that credible intervals for p_j are wider at the ends of the fitted sequence. We recommend taking N at least as large as $2k$. In our view, the credible intervals for the shorter sequences (that are consistent with each other) are generally the better estimates. In simulation studies (not shown), we found

1. no evidence of bias in point estimates of k;
2. for large k (≈ 100), the 95% credible intervals contained k in 98% (sd $= 0.2\%$) of simulations;

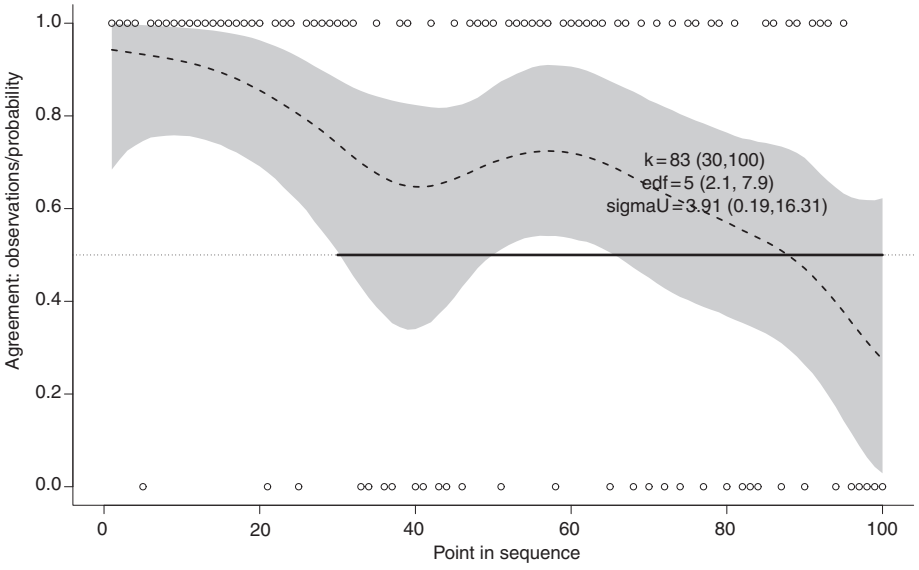

Figure 3: Same as Figure 1 with $N = 100$, $\delta = 100$, 21 df.

3. for small k (≈ 10), the 95% credible intervals contained k in 92% (sd $= 1.7\%$) of simulations.

Thus, for large k, the 95% credible intervals would seem to have greater than 95% coverage, and for small k, a little less than 95% coverage.

Tables 2–5 show the 95% credible interval may be sensitive to the number (N) of agreement points fitted. When using our program, we would suggest fitting the full agreement sequence using 20 or so degrees of freedom, and then refitting using a shorter sequence, taking N to be greater than the number of objects which may be wanted for further processing, while allowing roughly the same number of $\hat{p}_j$ above, as below 0.5, and choosing the degrees of freedom to allow at least five points between knots.

In general, the TopKlists estimates for k lie within the credible intervals given by our method. However, when δ is large (100) TopKLists estimates can vary considerably. Thus, in Table 2, estimates for k take values from 34 to 78. Firstly it probably makes sense to match δ in some way with v, and secondly, the more consistent results are those to be believed.

Standard diagnostics for MCMC runs are produced by the program and show satisfactory behaviour.

Tables 2–5 show that estimates for k and its credible intervals show very little sensitivity to the choice of the number of degrees of freedom. Estimates for k show some sensitivity to the choice of sequence length fitted (Table 3), but in general these differences are relatively minor.

Comparing estimates for k for each pairing of ranked data in Tables 2–5 we see clearly that the rankings for MDCC and Pusztai are more closely aligned than the other two pairings. This conclusion can be drawn more comfortably when credible intervals are given, less comfortably when only point estimates are available.

(Note that estimates for k and its credible intervals can differ by 1 for different MCMC runs, leading to occasional discrepancies between tables and figures.)

4 Discussion

H-S use several parameters in implementing their algorithm: δ, the distance allowed for ranks to be considered to be in agreement; v, the window for finding estimates for p_j; and $C(> .25)$, which controls the 'moderate deviations' of the probability, p_j (Rubin and Sethuraman, 1965). H-S use a locally defined value of $\hat{p}_j$ (based on the window, v) to determine the value k at which p_j changes from being above 0.5 to below 0.5. When using the TopKLists algorithm one must choose values for each of these. Generally, having set δ, k tends not to be

Table 3: Using different sequence lengths (N) for the breast cancer data: TransBig vs. MDCC.

δ	Method	N	df	v	k	95% CI
6	Penalized spline	917	15		0	(0, 18)
			21		0	(0, 19)
			27		0	(0, 19)
		200	11		11	(0, 21)
			21		11	(0, 20)
		150	31		12	(0, 21)
		100	11		12	(0, 20)
			21		12	(0, 21)
		50	11		12	(0, 20)
			19		12	(0, 20)
10	Penalized spline	917	27		0	(0, 22)
		400	21		10	(0, 25)
		200	21		11	(0, 24)
		150	31		12	(0, 24)
		100	21		13	(0, 23)
		50	13		13	(0, 23)
20	Penalized spline	917	43		7	(0, 29)
		400	21		15	(0, 31)
		200	21		16	(0, 31)
		150	31		17	(0, 31)
		100	21		18	(0, 31)
		50	13		17	(0, 31)
30	Penalized spline	917	43		17	(0, 39)
		400	21		22	(0, 41)
		200	21		21	(0, 42)
		150	31		23	(0, 43)
		100	21		22	(0, 43)
		50	13		19	(0, 41)
40	Penalized spline	917	27		37	(4, 56)
		917	43		37	(3, 56)
		400	21		39	(14, 59)
		200	21		38	(15, 62)
		150	31		36	(15, 63)
		100	21		34	(14, 66)
100	Penalized spline	917	27		96	(72, 124)
		917	117		97	(72, 123)
		400	21		91	(67, 121)
		200	21		91	(65, 121)
		150	23		88	(37, 120)
		100	21		83	(31, 100)

very sensitive to the choice of C or v. However, major disadvantages are the lack of an interval for k, and the lack of visualisation of the probability of agreement.

In choosing the number of degrees of freedom for the spline smooth, the number needs to be sufficiently large to fit the data (Ruppert, 2002). Marley and Wand (2010) suggest that 27 or so degrees of freedom may be sufficient.

Certainly we see very little difference in the fits and the estimation of k and its credible intervals, whether we choose 19 or 117 degrees of freedom. Estimates of k and its credible intervals are remarkably stable. However, in searching for k we are interested in the behaviour of the curve close to its start, rather than across the full agreement sequence, and using equally spaced knots with 51 degrees of freedom for a sequence of length 917 meant that each knot interval used approximately 18 data points, which may have forced too much smoothing.

Table 4: Using different sequence lengths (N): breast cancer data: MDCC vs. Pusztai.

δ	Method	N	df	v	k	95% CI
6	Penalized spline	917	17		11	(0, 23)
			21		11	(0, 23)
			27		11	(0, 23)
		150	19		18	(10, 25)
	TopKLists			10	14	
10	Penalized spline	917	19		10	(0, 24)
			27		10	(0, 24)
			27		10	(0, 25)
		150	19		19	(10, 27)
	TopKLists			30	15	
20	Penalized spline	917	21		10	(0, 24)
			27		10	(0, 25)
			51		22	(0, 37)
		150	19		26	(17, 37)
	TopKLists			20	20	
30	Penalized spline	917	21		38	(19, 52)
			51		38	(19, 52)
		150	19		35	(24, 51)
	TopKLists			20	24	
				40	28	
40	Penalized spline	917	17		52	(35, 67)
			27		52	(35, 67)
		150	19		43	(27, 71)
	TopKLists			30	30	
				40	32	
100	Penalized spline	917	15		120	(91, 155)
			21		123	(94, 159)
			50		124	(94, 159)
		300	51		109	(77, 161)
		200	19		96	(90, 162)
		150	19		92	(41, 144)
	TopKLists			30	87	
				70	92	

Searching for the point where the probabilities of agreement change from $p \geq 0.5$ to $p < 0.5$ is a local problem. Hence global summaries of the fit are not really appropriate, since (in the case of the splines) they may well be measuring how well we fit the bulk of the data when, in fact, the interest is in the fit for perhaps the first 200 observed agreements. The local fit issue means that we could well change the spacing of our knots from being equally distributed across the observations, or vary H-S's window, v, across the sequence. However, we prefer an approach which does not require precise tailoring of parameters to the data. Hence, our choice of knots at equal quantiles of the sequence, and the decision to look at the sensitivity of the conclusions to the choice of degrees of freedom for the spline smooth. One issue might be thought to be that the point, k, at which p_j changes from being above 0.5 to being below 0.5 may well occur in the section of the curve between the first (boundary) knot and the first internal knot. However, generally, any reasonable value of k, even between the lower boundary and the first knot, is sufficiently far away from the first point in the sequence, that the boundary behaviour of the spline is not an issue. When k is estimated as zero, the curve fits are convincing and not a function of boundary behaviour.

We note that dependent on the data and the purpose of the analysis, one may wish to find the point where the probabilities change from being above to being below a value that is different from 0.5. Our approach can be extended in a straightforward way to such a situation, but this is beyond the scope of this article.

Table 5: Using different sequence lengths (N): breast cancer data: TransBig vs. Pusztai.

δ	Method	N	df	v	k	95% CI
6	Penalized spline	917	15		8	(0, 18)
			27		8	(0, 18)
		200	31		14	(4, 21)
		150	19		14	(2, 21)
	TopKLists			10	10	
10	Penalized spline	917	15		7	(0, 18)
			27		7	(0, 18)
			51		7	(0, 18)
		200	31		14	(0, 22)
		150	19		14	(2, 22)
	TopKlists			30	11	
40	Penalized spline	917	15		25	(0, 41)
			21		25	(0, 41)
			51		29	(5, 43)
		200	31		31	(19, 43)
		150	19		32	(19, 44)
	TopKLists			30	24	
100	Penalized spline	917	15		51	(0, 80)
			30		51	(0, 81)
		300	31		47	(29, 72)
		200	25		45	(28, 68)
		150	19		43	(25, 68)
	TopKLists			30	28	
				70	35	

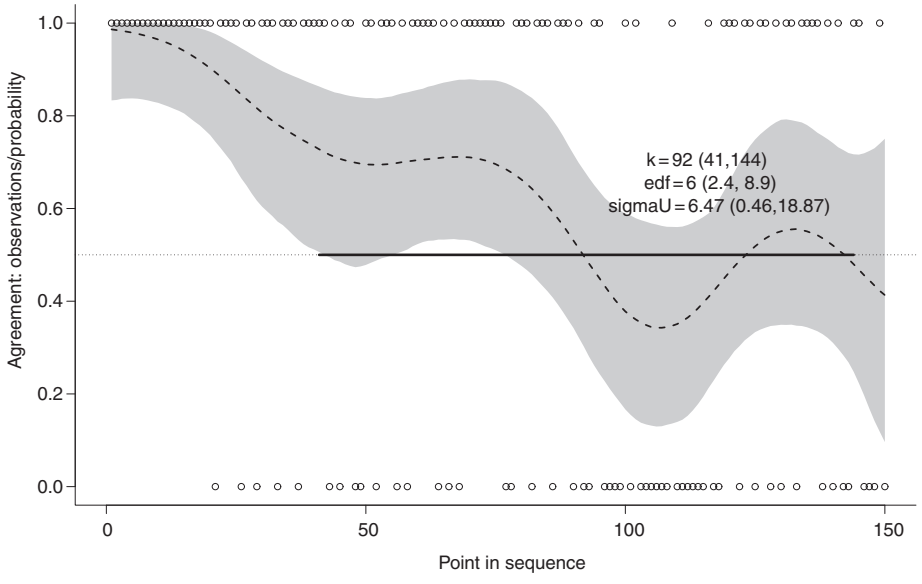

Figure 4: MDCC vs. Pusztai rankings: Otherwise same as Figure 2 with $N = 150$, $\delta = 100$, 19 df.

Wand and Ormerod (2008) illustrate the differences between smoothing splines, penalized splines and O'Sullivan splines for a number of examples, and show (for their examples) that the behaviour of the O'Sullivan splines is very close to that of smoothing splines at the boundaries. However, for ranked list comparisons, there are always several predictor points between the boundary knot and the first internal knot, so the boundary behaviours illustrated in Wand and Ormerod (2008) are less relevant. In any case, this uncertainty at the

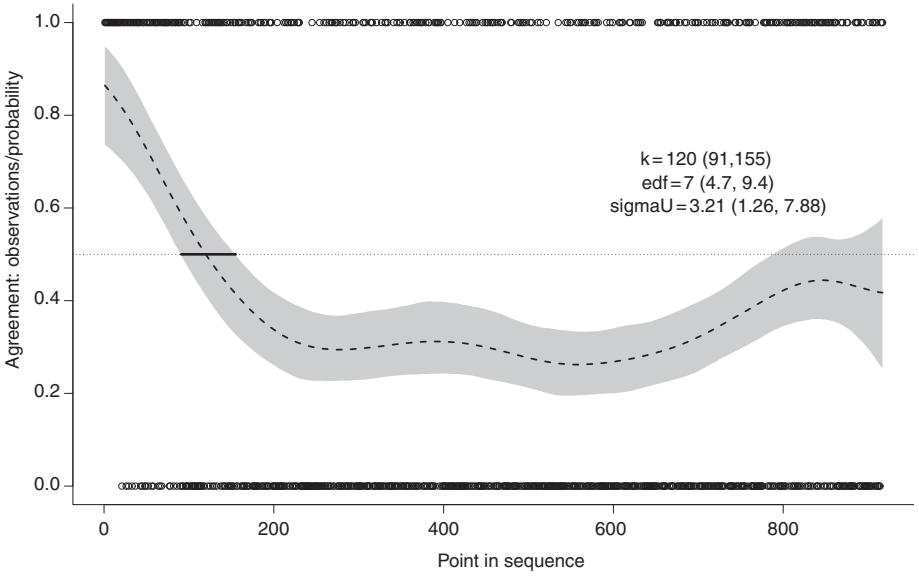

Figure 5: Same as Figure 4 with $N = 917$, $\delta = 100$, 15 df.

boundaries is effectively captured by the posterior distributions for p_j near the boundaries. This issue can only arise if the k is very small, in which case, it is largely irrelevant. We assessed sensitivity of the smoother by varying the number of degrees of freedom in the splines, verifying that the estimated degrees of freedom was lower than the fitted degrees of freedom, and looking for sensitivity in the estimates for k and its credible intervals.

Estimates for k and its credibility interval are sensitive to the fitted sequence length. See e.g. Table 3, when $\delta = 40$, where the lower bound (the 2.5% quantile of the posterior distribution) is 3 or 4 for the full sequence length but 14 or 15 for the shorter sequence lengths. We experimented by increasing the number of knots fitted to the full sequence so that there were the same number of observations between knots, as for a shorter sequence, expecting that the same fitted behaviour would then be observed. However, it is clear that σ_u^2 (and hence, the constant penalty, λ, estimated over the full sequence), dominates the fit, and lumps and bumps found in a shorter sequence are not found in the longer sequences no matter how many knots we fit (Figures 2 and 3, Table 3, $\delta = 100$.)

While it is generally the case that sequences of rankings of RNA-seq data, and of microarray data are of the order of tens of thousands, typically interest centers on the first 100–200 highest ranked genes. In such circumstances, it seems sensible to look for agreement of rankings within the first 1000–2000 genes. On occasion it may be the case, that a researcher is searching for the least differentially expressed genes. See, for example, Risso et al. (2014), where such genes are sought for normalization within an analysis. This can easily be done by reversing the rankings, and choosing the number of rankings which are of interest.

The webpage for TopKLists (Schimek et al., 2014) and the TopKLists vignette (Schimek et al., 2015a) suggest that the distance for agreement, δ, should be chosen using the 'deltaplot': "one should choose a value for δ, where the rate of the deltaplot's decrease begins to slow noticeably (i.e. where the discordance is starting to degrade)", and should there be several such points, the choice should be the smallest such value. (The deltaplot graphs the number of failures in agreement versus the distance, δ.) This makes sense, in that any of the earlier choices for δ should make little difference given a sharp decline. An upper limit to the choice of δ occurs when all the rankings are concordant (which occurs for random ranking agreement sequences when δ is chosen to be approximately $N/2$). This advice seems to be based on the fact that, typically, when the delta plot flattens out, TopKLists gives consistent values for k. However, given the many processes to which 'omics' data are subjected prior to finding a ranking, e.g. count normalisation, our view is that this choice should be left to the analyst, and be based on an understanding of the processes used to find the rankings.

A single estimate is returned by TopKLists and it can differ considerably for a given δ, depending on the window size, v. Further, a key problem is that, for a given δ and v, no insight into the variability of k is readily available. A strength of TopKLists is that it can deal with several lists of rankings, and that not all lists of rankings include the items contained by others. Our purpose was somewhat different: we sought to address the problem, as proposed by Hall and Schimek (2012), of determining when a set of paired rankings degenerates into noise, and to show that the problem had an existing solution in the context of nonparametric curve fitting, and to use that solution to find credible intervals for k, and to visualize the agreements more meaningfully.

5 Conclusions

To determine where a paired list of rankings degenerates into noise, a method based on the use of penalized splines and a Bayesian logit model has been developed. This approach produces a nonparametric smooth of the sequence of agreements between the two ranked lists and pairwise credible intervals for the probability of agreement, where agreement is defined as the two rankings differing by some fixed distance. Further, this method (a) gives a point estimate as well as a credible interval for the point at which agreement degenerates into noise and (b) enables the relationship between two rankings over any desired range to be evaluated. The initial approach to finding where a paired list of rankings degenerates into noise was developed by Hall and Schimek (2012) and gives a point estimate but no interval estimate making evaluation and comparison difficult. We have found that generally this point estimate lies within the credible interval given by our method.

Acknowledgment: This study was supported by the National Health and Medical Research Council (NHMRC grant # 525453). Comments by a reviewer substantially improved the paper.

Conflict of interest statement: None declared.

Appendix A: Technical details

A.1 Finding k, and its 95% credible intervals

The smoothing function, $p_j = 1/(1 + \exp(-\mu_j))$ where $\mu_j = \mathbf{X}\beta + \mathbf{Zu}$ and $j = 1, \ldots, N$, forms the basis for finding k and its credible interval. Let the sample of p_j at iteration t (of T samples) in the MCMC process be p_{j_t}. We monitor p_{j_t}, thus giving T posterior MCMC samples of p_{j_t} from which we estimate p_j and its 95% credible intervals. The posterior credible interval for k is found by post processing the posterior distributions p_{j_t}, to find k_t at each (post burn-in) iteration, t. Thus, to find k, for each iteration, we perform the following calculation:

1. If $p_{1t} \leq 0.5$, then $k_t = 0$; else
2. find the first j_t, for which $p_{j_t} - 0.5 < 0$, then $k_t = j - 1$; else
3. if there is no j_t, for which $p_{j_t} - 0.5 < 0$, then $k_t = N$.

That is, k_t is the last value of j before p_{j_t} becomes less than 0.5. See the post processing R-code below.

A fundamental output of the method is a figure showing the pointwise estimates of p_j (the median of p_j) and their 95% credible intervals. We graph the pointwise 95% credible intervals as a shaded area, and join the medians with a dashed line. A horizontal solid line shows the 95% credible interval for k, which does not necessarily correspond to the shaded interval at probability $= 0.5$, but is calculated from the posterior distribution of k.

A.2 Calculation of effective degrees of freedom and the spline penalty

When fitting splines, one needs to know whether sufficient knots have been fitted. Does the model adequately fit the data? The sensitivity of the model to the choice of the number of degrees of freedom can be checked by fitting models with various degrees of freedom. Marley and Wand (2010), in a different context, suggest that 25 degrees of freedom should be adequate. Here with sequences of length often greater than 1000, more knots may well be needed. (Ruppert, 2002; Wang et al., 2011) suggest that 40 knots may be adequate, however large N may be.)

We find the estimated degrees of freedom (*edf*) for the spline model fit as follows. Adapting Marley and Wand (2010); Wand (2014), the design matrix for penalized spline fit is $\mathbf{W} = [\mathbf{X}\ \mathbf{Z}]$, κ is the number of internal knots, and N is the length of the sequence. Let

$$\begin{aligned} \mathbf{D} &= diag(\mathbf{0}_2, \mathbf{1}_{\kappa+2}), \\ \Sigma_p &= diag(p_j(1-p_j)),\ j = 1, 2, \ldots N, \\ \text{and } p_j &= 1/(1+\exp(-\mu_j)). \end{aligned}$$

Then the effective degrees of freedom, *edf*, derived from Ruppert et al. (2003), Marley and Wand (2010) is given by:

$$\begin{aligned} &edf = \text{trace}[(\mathbf{W}^T\Sigma_p\mathbf{W} + \lambda^2\mathbf{D})^{-1}(\mathbf{W}^T\Sigma_p\mathbf{W})], \text{ where} \\ &\lambda^2 = 1/\sigma_u^2, \text{ and} \\ &\lambda \text{ is the spline penalty.} \end{aligned}$$

The *edf*, σ_u^2, λ and all other parameters in the model, are found from their posterior distributions. They are not chosen *a-priori*, nor are they arbitrary. They approximate maximum likelihood solutions for the given model, when N is large in comparison with the number of fitted parameters.

When the fitted degrees of freedom well exceeds the estimated degrees of freedom, it is clear that the smoothing has not been constrained by the choice of the number of fitted degrees of freedom.

BUGS code for penalized spline logit model

```
model{
 for (j in 1:N){
           mu[j] <- b + b1*x[j] + inprod(a[],zz[j,])
           logit(p[j]) <- mu[j]
           y[j] ~ dbern(p[j])
     }

 for (m in 1:numKnots){
      a[m] ~ dnorm(0,tauU)
     }
   b ~ dnorm(0, .0001); b_1 ~ dnorm(0, .0001)
   numerU ~ dnorm(0,1) ; denomU ~ dnorm(0,0.0016)
   tauU <- pow(numerU/denomU,2)
}
```

The code is valid in WinBUGS, OpenBUGS or JAGS.

The model uses a Half-Cauchy prior for the variance of the coefficients for the penalized spline terms. Marley and Wand (2010) recommend these and in our experience, they work well, allowing the model to initialize and adapt with no problems, while being uninformative. Code to generate the O'Sullivan splines over the number of knots for the predictor x is available from Wand and Ormerod (2008). When there are κ internal knots the fitted number of degrees of freedom (*df*) is $\kappa + 4$.

The last two lines of code show the technique for producing a Half-Cauchy(A) prior for σ_u^2, where $A = \text{sqrt}(1/.0016) = 25$.
The code is valid for OpenBUGS, WinBUGS, and JAGS.

A.3 R code for calculating *k* and its 95% credible interval

Let $\mathbf{P} = p_{j_t}$ be the matrix of posterior distributions for the p_j. The dimensions of this matrix are the number of iterations, T, and N, the length of the agreement data. Then the code for post processing $\mathbf{P}$ to find k is given by the function kCI($\mathbf{P}$):

```
kCI <- function(P){
     T <- dim(P)[1]; N <- dim(P)[2]
     k <-1:T
     for (t in 1:T){
         if (P[t,1]<=0.5) {k[t] <- 0 }
         else {for (j in 2:N){
              if ((P[t,j]-.5)<0) {k[t] <- j-1; break}
              else k[t] <-N}
   }}
   kCI <- as.numeric(quantile(k, probs=c(0.5, .025, .975)))
   return(kCI)
}
kCI(P)
```

Thus, we search through the iterations to find the first j for which P_{tj} is less then 0.5. The tth sample of k, k_t, is given by the value $j - 1$. At this point we exit the inner loop and return to the outer loop to find the next sample of k.

Note that if P_{t1} is less than or equal to 0.5 in iteration, t, then k_t is taken to be 0. If P_{tN} is greater than 0.5 at iteration t, when $j = N$, then k_t is taken to be N.

A.4 Further implementation details

For any given comparison, the agreement sequence is calculated using the TopKLists (Schimek et al., 2015b) function prepare.idata, which takes the two ranked lists and the desired distance, δ, to give the agreement sequence. The TopKLists function is used to ensure that agreement sequences are identical thereby allowing comparison with the estimated k from TopKLists. The spline bases are calculated via R code (ZOSull.R) supplied in Marley and Wand (2010). A function, wrapper(), outputs k and its credible intervals, together with p_j and its summary statistics, and various graphs to a nominated subdirectory. Input parameters are the set of ranked data (in the form used by TopKLists), the two columns to be compared, and the distance (δ) desired for defining agreement of ranks.

The models were fit using R2jags (Su and Yajima, 2015) which uses JAGS (Plummer, 2011). Models were also fit using BRugs (Thomas et al., 2006). However, the R2jags framework was found to be both considerably faster, and more flexible in that it allowed easy random initializations of several chains.

In JAGS, burn-in was 10,000, and the number of posterior MCMC iterations for estimation was 15,000. Autocorrelation graphs, mixing graphs and distribution plots for various parameters were plotted and found to be satisfactory. The models showed good mixing and autocorrelation properties.

Appendix B: Supplementary files

1. R code Vignette: RankAgreeVignetteV10.pdf: A vignette outlining the code needed to run the wrapper function and necessary ancillary code.
2. VignetteScriptV4.R: R code for the vignette.
3. functions_kpV10.R: The code for the wrapper function and other necessary functions. The code for ZOSull.R which generates the O'Sullivan spline bases is found at http://www.jstatsoft.org/article/view/v037i05 (Marley and Wand, 2010).
4. MCMCSupportingFunctions.R: Supporting functions which take the MCMC objects from jags and graph the estimated probability function, together with some diagnostic graphs and csv files summarising the probability functions.

References

Antosh, M., D. Fox, L. N. Cooper and N. Neretti (2013): "CORaL: comparison of ranked lists for analysis of gene expression data," J. Comput. Biol., 20, 433–443. http://dx.doi.org/10.1089/cmb.2013.0017.

Crainiceanu, C. M., D. Ruppert and M. P. Wand (2005): "Bayesian analysis for penalized spline regression using WinBUGS," J. Stat. Softw., 14, 1–24. http://www.jstatsoft.org/v14/i14/paper.

Desmedt, C., F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M. S. d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A. L. Harris, J. G. Klijn, J. A. Foekens, F. Cardoso, M. J. Piccart, M. Buyse and C. Sotiriou (2007): "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the transbig multicenter independent validation series," Clin. Cancer Res., 13, 3207–3214.

Dobson, A. J. and A. G. Barnett (2008): An introduction to generalized linear models, Chapman & Hall/CRC Texts in statistical science series, vol. 77, Boca Raton: CRC Press, 3rd edition.

Eden, E., D. Lipson, S. Yogev and Z. Yakhini (2007): "Discovering motifs in ranked lists of DNA sequences," PLoS Comput. Biol., 3, e39, http://dx.plos.org/10.1371.

Eden, E., R. Navon, I. Steinfeld, D. Lipson and Z. Yakhini (2009): "GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists," BMC Bioinformatics, 10, 48, http://www.biomedcentral.com/1471-2105/10/48.

Hall, P. and M. G. Schimek (2012): "Moderate-deviation-based inference for random degeneration in paired rank lists," J. Am. Stat. Assoc., 107, 661–672.

Hastie, T. and R. Tibshirani (1990): Generalized additive models, Monographs on statistics and applied probability, London, New York: Chapman and Hall, 1st edition.

Lottaz, C., X. Yang, S. Scheid and R. Spang (2006): "Orderedlist - a Bioconductor package for detecting similarity in ordered gene lists," Bioinformatics, 22, 2315–2316, http://bioinformatics.oxfordjournals.org/content/22/18/2315.abstract.

Lunn, D., C. Jackson, N. Best, A. Thomas and D. Spiegelhalter (2013): The BUGS book: a practical introduction to Bayesian analysis, Texts in statistical science, Boca Raton, FL: CRC Press.

Mallows, C. L. (1957): "Non-null ranking models. I," Biometrika, 44, 114–130, http://www.jstor.org/stable/2333244.

MAQC Consortium (2006): "The microarray quality control (MAQC): project shows inter- and intraplatform reproducibility of gene expression measurements," Nat. Biotechnol., 24, 1151 – 1161, http://www.nature.com/nbt/journal/v24/n9/full/nbt1239.html.

Marley, J. K. and M. P. Wand (2010): "Non-standard semiparametric regression via BRugs," J. Stat. Softw., 37, 1–30, http://www.jstatsoft.org/article/view/v037i05, http://www.jstatsoft.org/article/view/v037i05.

McCullagh, P. and J. A. Nelder (1989): Generalized linear models, Monographs on statistics and applied probability, vol. 37, London, New York: Chapman and Hall, 2nd edition.

Mood, A. M. (1940): "The distribution theory of runs," Ann. Math. Stat., 11, 367–392, http://www.jstor.org/stable/2235718.

O'Sullivan, F. (1986): "A statistical perspective on ill-posed inverse problems," Stat. Sci., 1, 502–527, http://projecteuclid.org/euclid.ss/1177013525.

Pihur, V., S. Datta and S. Datta (2014): RankAggreg: weighted rank aggregation, http://CRAN.R-project.org/package=RankAggreg, r package version 0.5.

Pinheiro, J., D. Bates, S. DebRoy, D. Sarkar and R Core Team (2016): nlme: linear and nonlinear mixed effects models, http://CRAN.R-project.org/package=nlme, r package version 3.1-128.

Plaisier, S. B., R. Taschereau, J. A. Wong and T. G. Graeber (2010): "Rank-rank hypergeometric overlap: identification of statistically significant overlap between gene-expression signatures," Nucleic Acids Res., 38, e169, http://nar.oxfordjournals.org/content/38/17/e169.abstract.

Plummer, M. (2011): JAGS Version 3.1. 0 user manual, http://gentoo.mirrors.lug.ro/freebsd/distfiles/mcmc-jags/jags_user_manual.pdf.

Risso, D., J. Ngai, T. P. Speed and S. Dudoit (2014): "Normalization of RNA-seq data using factor analysis of control genes or samples," Nat. Biotechnol., 32, 896–902.

Rubin, H. and J. Sethuraman (1965): "Probabilities of moderate deviations," Sankhya Indian J. Stat. Ser. A (1961–2002), 27, 325–346.

Ruppert, D. (2002): "Selecting the number of knots for penalized splines," J. Comput. Graph. Stat., 11, 735–757.

Ruppert, D., M. P. Wand and R. J. Carroll (2003): *Semiparametric regression*, Cambridge series in statistical and probabilistic mathematics, Cambridge, UK: Cambridge University Press.

Schimek, M. G., E. Budinska, J. Ding, K. G. Kugler, V. Svendova and S. Lin (2015a): "TopKLists: analyzing multiple ranked lists," https://cran.r-project.org/web/packages/TopKLists/vignettes/TopKLists.pdf.

Schimek, M. G., E. Budinska, K. G. Kugler, V. Svendova, J. Ding and S. Lin (2014): "TopKLists show case for integrating miRNA measurements," http://topklists.r-forge.r-project.org/showcase_miRNA/topklists-miRNA.html, accessed: August 25, 2016.

Schimek, M. G., E. Budinska, K. G. Kugler, V. Svendova, J. Ding and S. Lin (2015b): "TopKLists: a comprehensive R package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists," Stat. Appl. Genet. Mol. Biol., 14, 311–316, https://www.degruyter.com/view/j/sagmb.2015.14.issue-3/sagmb-2014-0093/sagmb-2014-0093.xml.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski and T. Ideker (2003) "Cytoscape: a software environment for integrated models of biomolecular interaction networks," Genome Res., 13, 2498–2504.

Stevens, W. L. (1939): "Distribution of groups in a sequence of alternatives," Ann. Eugen., 9, 10–17, http://dx.doi.org/10.1111/j.1469-1809.1939.tb02193.x.

Su, Y.-S. and M. Yajima (2015): R2jags: using R to Run 'JAGS', http://CRAN.R-project.org/package=R2jags, R package version 0.5-6.

Tabchy, A., V. Valero, T. Vidaurre, A. Lluch, H. Gomez, M. Martin, Y. Qi, L. J. Barajas-Figueroa, E. Souchon and C. Coutant (2010): "Evaluation of a 30-gene paclitaxel, fluorouracil, doxorubicin, and cyclophosphamide chemotherapy response predictor in a multicenter randomized trial in breast cancer," Clin. Cancer Res., 16, 5351–5361.

Thomas, A., B. O'Hara, U. Ligges and S. Sturtz (2006): "Making BUGS open," R News, 6, 12–17, http://cran.r-project.org/doc/Rnews/.

Wand, M. P. (2009): "Semiparametric and graphical models," Aust. N. Z. J. Stat., 51, 9–41.

Wand, M. P. (2014): *Semiparametric regression (short course, UTS, Sydney)*, http://matt-wand.utsacademics.info/sprSC.html, July 11, 2014.

Wand, M. P. and J. T. Ormerod (2008): "On semiparametric regression with O'Sullivan penalized splines," Aust. N. Z. J. Stat., 50, 179–198.

Wang, X., J. Shen and D. Ruppert (2011): "On the asymptotics of penalized spline smoothing," Electron. J. Stat., 5, 1–17.

Supplemental Material: The online version of this article (DOI: 10.1515/sagmb-2016-0036) offers supplementary material, available to authorized users.