

A Statistical View of Ranking: Midway between Classification and Regression

Yoonkyung Lee*¹

Department of Statistics
The Ohio State University

*joint work with Kazuki Uematsu

June 4-6, 2014

Conference on Nonparametric Statistics for Big Data
and Celebration to Honor Grace Wahba
Madison, WI

¹Research supported in part by NSF


Ranking

- ▶ Aims to order a set of objects or instances reflecting their underlying utility, quality or relevance to queries.
- ▶ Has gained increasing attention in machine learning, collaborative filtering and information retrieval for website search and recommender systems.





This item: Spline Models for Observational Data by Grace Wahba

Customers Who Bought This Item Also Bought

			
 A Practical Guide to Splines (Applied ... Carl de Boor ★★★★☆ (4)	Spline Functions: Basic Theory (Cambridge Mathematical Library) Larry Schumaker	Semiparametric Regression (Cambridge ... ▶ David Ruppert ★★★★★ (1)	Empirical Processes in M-Estimation ... Sara A. van de Geer ★★★★☆ (1)

Googling Wahba without Grace



[Web](#) [Videos](#) [Images](#) [News](#) [Maps](#) [More](#) [Search tools](#)

About 285,000 results (0.19 seconds)

Grace Wahba Home Page - Department of Statistics
www.stat.wisc.edu/~wahba/wahba.html • University of Wisconsin-Madison •
Grace Wahba. Good words from the NY Times: "For Today's Graduate, Just One Word: Statistics" here . "Advertising Companies Fret Over a Digital Talent Gap" ...

Youssef Wahba - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Youssef_Wahba • Wikipedia •
Youssef Wahba Pasha (1852-1934) (يوسف باشا وهبة) Egyptian Prime Minister and jurist. Youssef Wahba was born in Cairo, Egypt in 1852 of a prominent Coptic ...


Grace Wahba - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Grace_Wahba • Wikipedia •
Grace Wahba (born August 3, 1934) is the I. J. Schoenberg Professor of Statistics at the University of Wisconsin-Madison. She is a pioneer in methods for ...

Wahba's problem - Wikipedia, the free encyclopedia
en.wikipedia.org/wiki/Wahba's_problem • Wikipedia •
In applied mathematics, Wahba's problem, first posed by Grace Wahba in 1965, seeks to find a rotation matrix (special orthogonal matrix) between two ...

Grace Wahba - Google Scholar Citations
scholar.google.com/citations?user=2IPABNoAAAAJ... • Google Scholar •
Professor of Statistics, University of Wisconsin-Madison - Verified email at stat.wisc.edu
G Wahba SIAM Journal on Numerical Analysis 14 (4), 651-667, 662, 1977. A least squares estimate of satellite attitude. G Wahba SIAM review 7 (3), 409-409 ...

Wahba | Facebook
<https://www.facebook.com/WahbaMusic> •
Wahba. 141 likes, if the Hulk had a favorite worship leader, his name would be Wahba.






Grace Wahba



Grace Wahba is the I. J. Schoenberg Professor of Statistics at the University of Wisconsin-Madison. She is a pioneer in methods for smoothing noisy data.
[Wikipedia](#)

Born: August 3, 1934 (age 79)
Education: Cornell University, Stanford University


People also search for [View 10+ more](#)

 Emanuel Parzen	 Vladimir Vapnik	 Robert Tibshirani	 Bradley Efron	 Jianqing Fan
---	---	--	--	---

[Feedback](#)

See results about

Youssef Wahba (Former Prime Minister of Egypt)
Born: 1852, Cairo, Egypt
Died: 1934



Data for Ranking

object₁

positive

object₂

negative

⋮

⋮

object_{*n*-1}

positive

object_{*n*}

negative

How to order objects so that positive cases are ranked higher than negative cases?

Main Questions

- ▶ How to rank?
- ▶ What evaluation (or loss) criteria to use for ranking?
- ▶ What is the best ranking function given a criterion?
- ▶ How is it related to the underlying probability distribution for data?
- ▶ How to learn a ranking function from data?

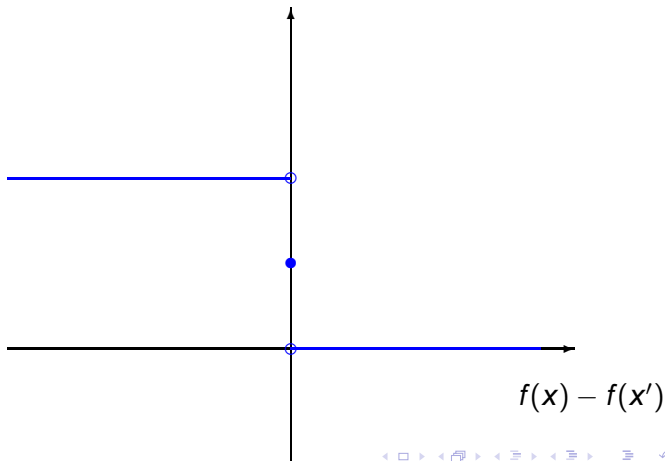
Notation

- ▶ $X \in \mathcal{X}$: an instance to rank
- ▶ $Y \in \mathcal{Y} = \{1, \dots, k\}$: an ordinal response in multipartite ranking (bipartite ranking when $k = 2$)
- ▶ $f: \mathcal{X} \rightarrow \mathbb{R}$: a real-valued ranking function whose scores induce ordering over the input space
- ▶ Training data: n pairs of (X, Y) from $\mathcal{X} \times \mathcal{Y}$

Pairwise Ranking Loss

For a pair of “positive” x and “negative” x' , define a loss of ranking function f as

$$\ell_0(f; x, x') = I(f(x) - f(x') < 0) + \frac{1}{2}I(f(x) - f(x') = 0)$$



Bipartite Ranking

- ▶ Note the invariance of the pairwise loss under order-preserving transformations.
- ▶ Find f minimizing the empirical ranking risk

$$R_{n_+, n_-}(f) = \frac{1}{n_+ n_-} \sum_{i=1}^{n_+} \sum_{j=1}^{n_-} \ell_0(f; \mathbf{x}_i, \mathbf{x}'_j)$$

- ▶ Minimizing ranking error is equivalent to maximizing AUC (area under ROC curve).

Likelihood Ratio Minimizes Ranking Risk

Clémenton et al. (2008), Uematsu and Lee (2011), and Gao and Zhou (2012)

Theorem

Define $f_0^*(x) = g_+(x)/g_-(x)$, and let $R_0(f) = E(\ell_0(f; X, X'))$ denote the ranking risk of f under the bipartite ranking loss. Then for any ranking function f ,

$$R_0(f_0^*) \leq R_0(f).$$

Remark

Connection to classification:

$$P(Y = 1|X = x) = \frac{\pi_+ g_+(x)}{\pi_+ g_+(x) + \pi_- g_-(x)} = \frac{f_0^*(x)}{f_0^*(x) + (\pi_-/\pi_+)}$$

Convex Surrogate Loss for Bipartite Ranking

- ▶ Exponential loss
in RankBoost (Freund et al. 2003):

$$\ell(f; \mathbf{x}, \mathbf{x}') = \exp(-(f(\mathbf{x}) - f(\mathbf{x}')))$$

- ▶ Hinge loss
in RankSVM (Joachims 2002) and AUCSVM
(Rakotomamonjy 2004, Brefeld and Scheffer 2005):

$$\ell(f; \mathbf{x}, \mathbf{x}') = (1 - (f(\mathbf{x}) - f(\mathbf{x}')))_+$$

- ▶ Logistic loss (cross entropy)
in RankNet (Burges et al. 2005):

$$\ell(f; \mathbf{x}, \mathbf{x}') = \log(1 + \exp(-(f(\mathbf{x}) - f(\mathbf{x}'))))$$

Optimal Ranking Function Under Convex Loss

Theorem

Suppose that ℓ is differentiable, $\ell'(s) < 0$ for all $s \in \mathbb{R}$, and $\ell'(-s)/\ell'(s) = \exp(s/\alpha)$ for some positive constant α .

Let f^ be the best ranking function f minimizing $R_\ell(f) = E[\ell(f; X, X')]$. Then*

$$f^*(x) = \alpha \log(g_+(x)/g_-(x)) \quad \text{up to a constant.}$$

Remark

- ▶ *For RankBoost, $\ell(s) = e^{-s}$, and $\ell'(-s)/\ell'(s) = e^{2s}$.
 $f^*(x) = \frac{1}{2} \log(g_+(x)/g_-(x))$.*
- ▶ *For RankNet, $\ell(s) = \log(1 + e^{-s})$, and $\ell'(-s)/\ell'(s) = e^s$.
 $f^*(x) = \log(g_+(x)/g_-(x))$.*

Ranking-Calibrated Loss

Theorem

Suppose that ℓ is convex, non-increasing, differentiable and $\ell'(0) < 0$. Then for almost every (x, z) , $\frac{g_+(x)}{g_-(x)} > \frac{g_+(z)}{g_-(z)}$ implies $f^(x) > f^*(z)$.*

Remark

For RankSVM, $\ell(s) = (1 - s)_+$ with singularity at $s = 1$ could yield ties in ranking (leading to inconsistency) while $\ell(s) = (1 - s)_+^2$ is ranking-calibrated.

RankSVM Can Produce Ties

Theorem

Let $f^ = \arg \min_f E(1 - (f(X) - f(X'))_+)$. Suppose that f^* is unique up to an additive constant.*

- (i) For discrete \mathcal{X} , a version of f^* is integer-valued.*
- (ii) For continuous \mathcal{X} , there exists an integer-valued function whose risk is arbitrarily close to the minimum risk.*

Remark

- ▶ *Scores from RankSVM exhibit granularity.*
- ▶ *Ranking with the hinge loss is not consistent!*

Extension to Multipartite Ranking

- ▶ In general ($k \geq 2$), for a pair of (x, y) and (x', y') with $y > y'$, define a loss of ranking function f as

$$\ell_0(f; x, x', y, y') = c_{y'y} I(f(x) < f(x')) + \frac{1}{2} c_{y'y} I(f(x) = f(x'))$$

where $c_{y'y}$ is the cost of misranking a pair of y and y' .
(*Waegeman et al. 2008*)

- ▶ Again, ℓ_0 is invariant under order-preserving transformations.

Optimal Ranking Function for Multipartite Ranking

Theorem

(i) When $k = 3$, let $f_0^*(x) = \frac{c_{12}P(Y = 2|x) + c_{13}P(Y = 3|x)}{c_{13}P(Y = 1|x) + c_{23}P(Y = 2|x)}$.

Then for any ranking function f ,

$$R_0(f_0^*; \mathbf{c}) \leq R_0(f; \mathbf{c}).$$

(ii) When $k > 3$ and let $f_0^*(x) = \frac{\sum_{i=2}^k c_{1i}P(Y = i|x)}{\sum_{j=1}^{k-1} c_{jK}P(Y = j|x)}$.

If $c_{1k}c_{ji} = c_{1i}c_{jk} - c_{1j}c_{ik}$ for all $1 < j < i < k$, then for any ranking function f ,

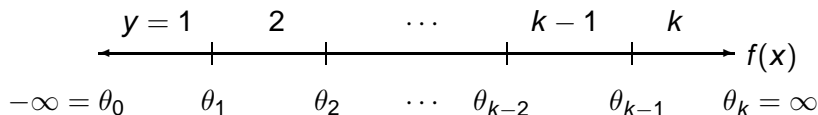
$$R_0(f_0^*; \mathbf{c}) \leq R_0(f; \mathbf{c}).$$

Remark

Let $c_{ji} = (s_i - s_j)w_i w_j I(i > j)$ for some increasing scale $\{s_j\}_{j=1}^k$ and non-negative weight $\{w_j\}_{j=1}^k$. e.g. $c_{ji} = (i - j)I(i > j)$

Ordinal Regression

- ▶ Ordinal regression is commonly used to analyze data with ordinal response in practice.



- ▶ A typical form of loss in ordinal regression for f with thresholds $\{\theta_j\}_{j=1}^{k-1}$:

$$\ell(f, \{\theta_j\}_{j=1}^{k-1}; \mathbf{x}, y) = \ell(f(\mathbf{x}) - \theta_{y-1}) + \ell(\theta_y - f(\mathbf{x})),$$

where $\theta_0 = -\infty$ and $\theta_k = \infty$.

Convex Loss in Ordinal Regression

- ▶ ORBoost (*Lin and Li 2006*):

$$\ell(s) = \exp(-s)$$

- ▶ Proportional Odds model (*McCullagh 1980, Rennie 2006*):

$$\ell(s) = \log(1 + \exp(-s))$$

- ▶ Support Vector Ordinal Regression (*Herbrich et al. 2000*):

$$\ell(s) = (1 - s)_+$$

Ordinal Regression Boosting (ORBoost)

- ▶ The optimal ranking function f^* under $\ell(s) = \exp(-s)$ is

$$f^*(x) = \frac{1}{2} \log \frac{\sum_{i=2}^k P(Y = i|x) \exp(\theta_{i-1}^*)}{\sum_{j=1}^{k-1} P(Y = j|x) \exp(-\theta_j^*)}$$

where θ_j^* are constants depending only on $P_{X,Y}$.

- ▶ When $k = 3$,

$$f^*(x) = \frac{1}{2} \log \frac{P(Y = 2|x) + \exp(\theta_2^* - \theta_1^*)P(Y = 3|x)}{\exp(\theta_2^* - \theta_1^*)P(Y = 1|x) + P(Y = 2|x)}$$

up to a constant. Hence, f^* preserves the ordering of f_0^* with $c_{12} = c_{23} = 1$ and $c_{13} = e^{\theta_2^* - \theta_1^*}$.

Proportional Odds Model

- ▶ Cumulative logits (McCullagh 1980)

$$\log \frac{P(Y \leq j|x)}{P(Y > j|x)} = f(x) - \theta_j,$$

where $-\infty = \theta_0 < \theta_1 < \dots < \theta_{k-1} < \theta_k = \infty$.

- ▶ Given $\{\theta_j\}_{j=1}^{k-1}$, maximizing the log likelihood amounts to ordinal regression with $\ell(s) = \log(1 + \exp(-s))$.
- ▶ When $k = 3$, given θ_1 and θ_2 , the minimizer of the deviance risk f^* satisfies

$$\exp(f^*(x)) = \frac{r(x) - 1 + \sqrt{(r(x) - 1)^2 + 4 \exp(\theta_1 - \theta_2) r(x)}}{2 \exp(-\theta_2)},$$

where $r(x) = \frac{P(Y = 2|x) + P(Y = 3|x)}{P(Y = 1|x) + P(Y = 2|x)} = f_0^*(x)$ with $c_{12} = c_{23} = c_{13} = 1$.

- ▶ When $\theta_2 > \theta_1$, $f^*(x)$ preserves the ordering of $r(x)$.

Support Vector Ordinal Regression

- ▶ SVOR with Implicit constraints in *Chu and Keerthi* (2007)

$$\ell(f, \{\theta_j\}_{j=1}^{k-1}; \mathbf{x}, y) = \sum_{j=1}^{y-1} (1 - (f(\mathbf{x}) - \theta_j))_+ + \sum_{j=y}^{k-1} (1 - (\theta_j - f(\mathbf{x})))_+.$$

- ▶ When $k = 3$, $f^*(\mathbf{x})$ is a **step function of**

$$r(\mathbf{x}) = \frac{p_2(\mathbf{x}) + p_3(\mathbf{x})}{p_1(\mathbf{x}) + p_2(\mathbf{x})} \text{ (i.e. } f_0^* \text{ with } c_{12} = c_{13} = c_{23}).$$

$r(\mathbf{x})$	$(0, \frac{1}{2})$	$(\frac{1}{2}, 1)$	$(1, 2)$	$(2, \infty)$
$f^*(\mathbf{x})$	$\theta_1 - 1$	$\min(\theta_1 + 1, \theta_2 - 1)$	$\max(\theta_1 + 1, \theta_2 - 1)$	$\theta_2 + 1$

Numerical Illustration

- ▶ Simulation setting:

$X|Y = 1 \sim N(-2, 1)$, $X|Y = 2 \sim N(0, 1)$ and
 $X|Y = 3 \sim N(2, 1)$

- ▶ When $c_{12} = c_{23} = c_{13} = 1$,

$$f_0^*(x) = \frac{P(Y = 2|X = x) + P(Y = 3|X = x)}{P(Y = 1|X = x) + P(Y = 2|X = x)} = \frac{e^{2x} + e^2}{e^{-2x} + e^2}.$$

- ▶ Generate 500 observations in each category.
- ▶ Apply pairwise ranking risk minimization with exponential loss, proportional odds model, ORBoost and SVOR.

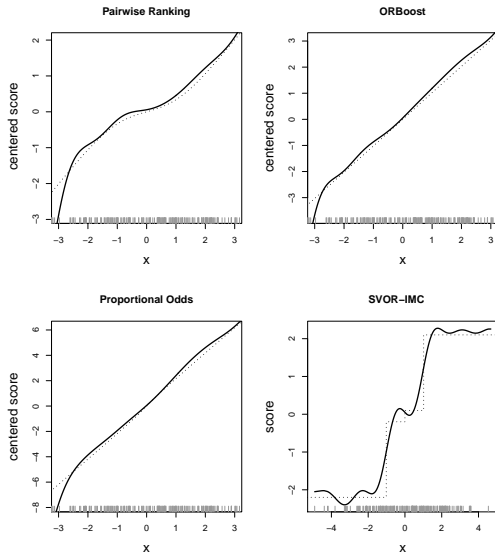


Figure: Theoretical ranking function (dotted line) and estimated ranking function (solid line) for pairwise ranking risk minimization with exponential loss, ORBoost, proportional odds model and SVOR with implicit constraints.

Application to Movie-Lens Data

- ▶ The data set consists of 100,000 ratings (on a scale of 1 to 5) for 1,682 movies by 943 users (GroupLens-Research).
- ▶ Contains content information about the movies (release date and genres) and demographic information about the users (age, gender and occupation).
- ▶ Transform five categories into three categories: “Low” (1-3), “Middle” (4) and “High” (5) and check the analytical results in $k = 3$.

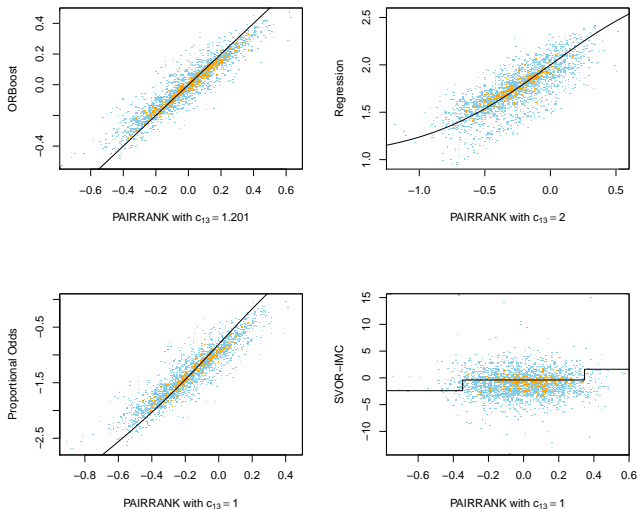


Figure: Scatter plots of ranking scores from ORBoost, regression, proportional odds model, and SVOR against pairwise ranking scores with matching cost c_{13} for MovieLens data with three categories. The solid lines indicate theoretical relation between ranking scores.

Concluding Remarks

- ▶ Provide a statistical view of ranking by identifying the optimal ranking function given loss criteria
- ▶ For pairwise multipartite ranking, the optimal ranking depends on the ratio of conditional probability weighted by misranking costs.
- ▶ Illustrate the connection between ranking and classification/ordinal regression in the framework of convex risk minimization.
- ▶ Our study bridges traditional methods such as proportional odds model in statistics with ranking algorithms in machine learning.

Special Thanks to Grace








*Behold the sower in the field,
With her arm she scatters the seeds.
Some seeds are trodden in the pathway;
Some seeds fall on stony ground.*

*But some seeds fall on fallow ground
They grow and multiply a thousand fold.*

– From Pete Seeger's "Sower Of Seeds"

References

-  S. Agarwal, and P. Niyogi. (2009) “Generalization Bounds for Ranking Algorithms via Algorithmic Stability.” *Journal of Machine Learning Research*, 10, 441- 474.
-  W. Chu, and S.S. Keerthi. (2007) “Support vector ordinal regression.” *Neural computation*, 19(3), 792- 815.
-  S. Cléménçon, G. Lugosi and N. Vayatis. (2008) “Ranking and empirical minimization of U-statistics.” *Annals of Statistics*, 36, 844- 874.
-  W. Gao and Z. Zhou (2012) “On the Consistency of AUC Optimization” *ArXiv:1208.0645*
-  R. Herbrich, T. Graepel, and K. Obermayer. (2000) “Large margin rank boundaries for ordinal regression.” *Advances in Large Margin Classifiers*, 115- 132.
-  P. Li, C. Burges, and Q. Wu. (2007) “McRank: Learning to Rank Using Multiple Classification and Gradient Boosting” *Advances in Neural Information Processing Systems*, 20, 897- 904.

-  H. Lin, and L. Li. (2006) "Large-Margin Thresholded Ensembles for Ordinal Regression: Theory and Practice." *Lecture Notes in Computer Science*, 4264, 319- 333.
-  P. McCullagh. (1980) "Regression Models for Ordinal Data." *Journal of Royal Statistical Society (B)*, 42(2), 109- 142.
-  J.D.M Rennie. (2006) "A comparison of McCullagh's proportional odds model to modern ordinal regression algorithms."
-  A. Shashua, and A. Levin. (2003) "Ranking with large margin principle: Two approaches." *Advances in Neural Information Processing Systems*, 15, 937- 944.
-  K. Uematsu, and Y. Lee. (2011) "On Theoretically Optimal Ranking Functions in Bipartite Ranking." *Technical Report No. 863, Department of Statistics, The Ohio State University.*
-  K. Uematsu, and Y. Lee. (2013) "Statistical Optimality in Multipartite Ranking and Ordinal Regression" *Technical Report No. 873, Department of Statistics, The Ohio State University.*
-  W. Waegeman, B. D. Baets and L.Boullart. (2008) "ROC analysis in ordinal regression learning." *Pattern Recognition Letters*, 29(1), 1- 9.