Stat Papers (2009) 50:887–893
DOI 10.1007/s00362-009-0263-3

Rank correlation estimators and their limiting distributions

Wojciech Niemiro · Wojciech Rejchel

Received: 11 September 2008 / Accepted: 15 May 2009 / Published online: 5 August 2009
© Springer-Verlag 2009

Abstract We examine a new rank correlation estimator, recently proposed by Bobrowski (Ranked modelling of risk on the basis of survival data. ICSMRA, Lisbon, 2007). It is obtained by minimization of a convex piece-wise linear criterion function. The main advantage of this estimator is the fact that it can be effectively computed by algorithms related to linear programming. We prove basic asymptotic theorems about the estimator: consistency and asymptotic normality.

Keywords Ranking · Linear ranking rule · Discontinuous criterion function · Support vector machines · Convex minimization · U-statistics

1 Introduction

1.1 The problem of ranking and the MRC estimator

The goal of ranking is to predict the order between objects (instances) on the basis of their observed features. We consider a population of objects equipped with a relation of (linear) ordering. For any two distinct objects o_1 and o_2 it holds either $o_1 \preceq o_2$ or $o_1 \succeq o_2$ (or maybe both), but it is unknown which is true. We lose little generality by

W. Niemiro (✉) · W. Rejchel
Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruń, Poland
e-mail: wniem@mat.uni.torun.pl

W. Rejchel
e-mail: iggypop@mat.uni.torun.pl

W. Niemiro
Institute of Applied Mathematics and Mechanics, University of Warsaw, Warsaw, Poland
e-mail: wniem@mimuw.edu.pl

assuming that real numbers y_1 and y_2 are assigned to the objects o_1 and o_2 in such a way that $o_1 \preceq o_2$ is equivalent to $y_1 \leq y_2$. However it is important that statistical procedures should use only linear ordering of the real line, that is they have to be invariant with respect to increasing transformations of the variables y_i. Let d-dimensional vectors $\mathbf{x_1}$ and $\mathbf{x_2}$ describe observed or measured features of the objects. We are to construct a function $\phi : \mathcal{X} \to \mathbb{R}$, called ranking rule, which predicts the order between the objects in the following way:

$$\text{if } \ \phi(\mathbf{x_1}) \leq \phi(\mathbf{x_2}) \text{ then we predict that } \ y_1 \leq y_2.$$

To measure the quality of ranking rule ϕ, we introduce a probabilistic setting. Let us assume that two objects are randomly selected from the population. They are described by a pair of independent and identically distributed random vectors $\mathbf{Z_1} = (\mathbf{X_1}, Y_1)$ and $\mathbf{Z_2} = (\mathbf{X_2}, Y_2)$ taking values in $\mathcal{X} \times \mathbb{R}$. Here $\mathcal{X}$ (observation space) is a measurable subset of $\mathbb{R}^d$. Random vectors $\mathbf{X_i}$ are regarded as observations, while Y_i are unknown variables which define the ordering.

Most natural approach is to seek ϕ which minimizes the probability of incorrect ranking:

$$\mathbb{P}(Y_1 > Y_2, \phi(\mathbf{X_1}) \leq \phi(\mathbf{X_2})). \tag{1}$$

The model of ranking similar to that described above was introduced by Han (1987). It is relevant to important applications in survival analysis and other branches of applied statistics. In fact Han considered linear ranking rules $\phi(\mathbf{x}) = \boldsymbol{\theta}^T\mathbf{x}$, where $\boldsymbol{\theta} \in \mathbb{R}^d$. According to criterion (1) we are thus to find $\boldsymbol{\theta_0}$ which minimizes

$$\Gamma(\boldsymbol{\theta}) = \mathbb{P}(Y_1 > Y_2, \boldsymbol{\theta}^T\mathbf{X_1} \leq \boldsymbol{\theta}^T\mathbf{X_2}). \tag{2}$$

Assume that we have an access to a learning sample, that is independent, identically distributed random vectors $\mathbf{Z_1} = (\mathbf{X_1}, Y_1), \ldots, \mathbf{Z_n} = (\mathbf{X_n}, Y_n)$ for which the ordering of components Y_i is observable. Then we can consider a sample analog of (2), namely

$$\Gamma_n(\boldsymbol{\theta}) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}(Y_i > Y_j, \boldsymbol{\theta}^T\mathbf{X_i} \leq \boldsymbol{\theta}^T\mathbf{X_j}). \tag{3}$$

We can use $\boldsymbol{\theta_n}$, the minimizer of $\Gamma_n(\boldsymbol{\theta})$, as an estimator of unknown parameter $\boldsymbol{\theta_0}$, the minimizer of $\Gamma(\boldsymbol{\theta})$. Han (1987) obtained the estimate $\boldsymbol{\theta_n}$, in fact, by maximizing $-\Gamma_n(\boldsymbol{\theta})$ and called $\boldsymbol{\theta_n}$ the *maximum rank correlation* (MRC) estimator. Statistical properties of $\boldsymbol{\theta_n}$ are of obvious interest. Han (1987) showed that his MRC estimator is consistent. Asymptotic normality is examined in his later paper Han (1988). A simpler proof of asymptotic normality can be found in Sherman (1993).

Note that $\Gamma_n(\boldsymbol{\theta})$ is, for every fixed $\boldsymbol{\theta} \in \mathbb{R}^d$, a U-statistic of order two. Therefore the object of investigations is a U-process $\{\Gamma_n(\boldsymbol{\theta}); \boldsymbol{\theta} \in \mathbb{R}^d\}$. The chief difficulty is discontinuous nature of "sample functions" of this process. Not only asymptotic analysis is quite hard, but (which is more important) minimization of (3) creates serious computational problems. Abrevaya (1999) developed an improved algorithm for computing

the objective function (3), but really effective algorithms for *minimizing* this function are not available. Known facts about related optimization problems in classification theory (Bartlett and Ben-David 2002) even suggest that such algorithms are unlikely to exist. Anyway, the lack of computational efficiency is probably the main obstacle to wider use of MRC estimates in practice.

1.2 The ψ-MRC estimator

To overcome the computational problems, discontinuous objective function can be replaced by a convex function designed to serve a similar purpose. This trick has been very successfully used in classification and led to the breakthrough invention of support vector machines (SVM). Clémençon et al. (2008) transferred the ideas behind SVM to the setup of ranking. Bobrowski (2007) proposed minimizing a convex and piece-wise linear criterion function in the context of ranking problems. A ranking counterpart of this criterion, which will be referred to as ψ-function, is the main focus in our paper. It is defined by

$$\Psi_n(\boldsymbol{\theta}) = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}(Y_i > Y_j)\, \psi(\boldsymbol{\theta}^T \mathbf{X_i} - \boldsymbol{\theta}^T \mathbf{X_j}), \tag{4}$$

where $\psi(t) = \max(0, 1-t)$. This function is convex, nonnegative and piece-wise linear. There exist algorithms, which compute the minimizer of $\Psi_n(\boldsymbol{\theta})$ effectively. Bobrowski and Niemiro (1984) devised an algorithm for minimization of a similar, convex and piece-wise linear, criterion function used in classification problems. As noted by Bobrowski (2007), the same method can be used to minimize (4). The algorithm of Bobrowski and Niemiro is based on the fact that the minimizer is at a "vertex", that is an intersection of hyperplanes at which the objective function is not differentiable. The algorithm searches the set of vertices and finds a minimum in finite number of steps. Alternative way is to use one of standard methods of linear programming, for example the *simplex algorithm*. Bloomfield and Steiger (1983) explain how to use such algorithms to compute the "least absolute deviations" linear regression. After minor modifications the same method can be applied to minimizing (4).

With slight abuse of notation we denote the minimizer of $\Psi_n(\boldsymbol{\theta})$ by $\boldsymbol{\theta_n}$ and call it the ψ-MRC estimate. The theoretical counterpart of (4) is

$$\Psi(\boldsymbol{\theta}) = \mathbb{E}\, \mathbb{I}(Y_1 > Y_2)\, \psi(\boldsymbol{\theta}^T \mathbf{X_1} - \boldsymbol{\theta}^T \mathbf{X_2}).$$

We assume that a point $\boldsymbol{\theta_0}$ which minimizes $\Psi(\boldsymbol{\theta})$ exists and is unique. This assumption will not be repeated in further statements.

2 Strong consistency and asymptotic normality

Our aim is to show consistency and asymptotic normality of $\boldsymbol{\theta_n}$, regarded as an estimator of unknown parameter $\boldsymbol{\theta_0}$. In our analysis we do not need as sophisticated methods

as those used by Han (1988) or Sherman (1993), because our criterion function is convex. The methods of proofs in our paper are rather analogous to Haberman (1989) and Niemiro (1992, 1993) who considered minimization of empirical processes under the assumption of convexity, but here we need extension of these methods to U-processes. Related problems are considered also by Bose (1998) in a different context.

Since convexity plays crucial role in our proofs, we will formulate our main results in a slightly more general way and then specialize them to the case of ψ-MRC estimators. We thus consider a function $Q : \mathbb{R}^d \to \mathbb{R}$ given by

$$Q(\boldsymbol{\theta}) = \mathbb{E} f(\boldsymbol{\theta}, \mathbf{Z_1}, \mathbf{Z_2}).$$

where $f(\cdot, \mathbf{z_1}, \mathbf{z_2})$ is convex for all $\mathbf{z_1}, \mathbf{z_2}$ and its sample analogue,

$$Q_n(\boldsymbol{\theta}) = \frac{1}{n(n-1)} \sum_{i \neq j} f(\boldsymbol{\theta}, \mathbf{Z_i}, \mathbf{Z_j}),$$

where random vectors $\mathbf{Z_1}, \ldots, \mathbf{Z_n}$ are independent and identically distributed. Let $\boldsymbol{\theta_0}$ and $\boldsymbol{\theta_n}$ denote minimizers of $Q(\boldsymbol{\theta})$ and $Q_n(\boldsymbol{\theta})$, respectively. In the proofs we need the following lemma, taken from Niemiro (1992), which is an easy consequence of standard results on convex functions (Rockafellar 1970).

Lemma 1 *Let $h_n(\boldsymbol{\theta})$, $n = 1, \ldots$ be convex random functions on $\mathbb{R}^d$. Assume that $h(\boldsymbol{\theta})$ is a random function such that $h_n(\boldsymbol{\theta}) \to h(\boldsymbol{\theta})$ (a) in probability (b) almost surely, for each fixed $\boldsymbol{\theta}$. Then on each compact $K \subset \mathbb{R}^d$ we have uniform convergence:*

$$\sup_{\boldsymbol{\theta} \in K} |h_n(\boldsymbol{\theta}) - h(\boldsymbol{\theta})| \longrightarrow 0.$$

(a) in probability (b) almost surely, respectively.

Theorem 1 *If $\boldsymbol{\theta_0}$ is the unique minimizer of $Q(\boldsymbol{\theta})$ then $\boldsymbol{\theta_n}$ is a strongly consistent estimator of $\boldsymbol{\theta_0}$.*

Note that under the assumptions of Theorem 1, a point $\boldsymbol{\theta_n}$ which minimizes $Q_n(\boldsymbol{\theta})$ exists almost surely, at least for sufficiently large n. It may be not unique, but then we can choose $\boldsymbol{\theta_n}$ arbitrarily subject to condition that selection is measurable. These facts will be easily seen in the proof. The existence of a measurable selector can be shown just as in Niemiro (1992, Appendix).

If we consider function $f : \mathbb{R}^d \times \mathbb{R}^{d+1} \times \mathbb{R}^{d+1} \to \mathbb{R}$ given by the formula

$$f(\boldsymbol{\theta}, \mathbf{z_1}, \mathbf{z_2}) = \mathbb{I}(y_1 > y_2)\, \psi(\boldsymbol{\theta}^T \mathbf{x_1} - \boldsymbol{\theta}^T \mathbf{x_2}), \tag{5}$$

where $\mathbf{z_i} = (\mathbf{x_i}, y_i)$, $\mathbf{x_i} \in \mathbb{R}^d$, $y_i \in \mathbb{R}$ for $i = 1, 2$, then we obtain Ψ and Ψ_n as special cases of Q and Q_n. Therefore we have the following result:

Corollary 1 *The ψ-MRC estimator is strongly consistent.*

Proof (Proof of Theorem 1) Continuity of Q and the fact that $\boldsymbol{\theta_0}$ is its unique minimizer imply that for arbitrary $\varepsilon > 0$ there exists $\tau > 0$ such that $Q(\boldsymbol{\theta}) > Q(\boldsymbol{\theta_0}) + 2\tau$ for $|\boldsymbol{\theta} - \boldsymbol{\theta_0}| = \varepsilon$. SLLN for U-statistics and Lemma 1 implies $\sup_{|\boldsymbol{\theta}-\boldsymbol{\theta_0}|\leq\varepsilon} |Q_n(\boldsymbol{\theta}) - Q(\boldsymbol{\theta})| \longrightarrow_{as} 0$. Then the following inequalities hold with probability one for sufficiently large n:

$$Q_n(\boldsymbol{\theta}) > Q(\boldsymbol{\theta}) - \tau > Q(\boldsymbol{\theta_0}) + \tau \quad \text{for } |\boldsymbol{\theta} - \boldsymbol{\theta_0}| = \varepsilon,$$

$$Q_n(\boldsymbol{\theta_0}) < Q(\boldsymbol{\theta_0}) + \tau.$$

Summarizing, $Q_n(\boldsymbol{\theta}) > Q_n(\boldsymbol{\theta_0})$ for every $\boldsymbol{\theta}$ such that $|\boldsymbol{\theta} - \boldsymbol{\theta_0}| = \varepsilon$. By convexity of Q_n we get $|\boldsymbol{\theta_n} - \boldsymbol{\theta_0}| < \varepsilon$. □

Now we have to introduce a few further definitions that we need in the proof of asymptotic normality. Let $g(\boldsymbol{\theta}, \mathbf{z_1}, \mathbf{z_2})$ be a subgradient of convex function $f(\boldsymbol{\theta}, \mathbf{z_1}, \mathbf{z_2})$. We do not require that it is unique and actually it is not in the case when f is defined as (5). We only require that g is a measurable selection of subgradient and refer again to Niemiro (1992, Appendix) for details. From the definition of subgradient (see Rockafellar (1970)) we get the following useful inequality

$$\begin{aligned} 0 &\leq f(\boldsymbol{\theta}, \mathbf{z_1}, \mathbf{z_2}) - f(\mathbf{0}, \mathbf{z_1}, \mathbf{z_2}) - \boldsymbol{\theta}^T g(\mathbf{0}, \mathbf{z_1}, \mathbf{z_2}) \\ &\leq \boldsymbol{\theta}^T [g(\boldsymbol{\theta}, \mathbf{z_1}, \mathbf{z_2}) - g(\mathbf{0}, \mathbf{z_1}, \mathbf{z_2})]. \end{aligned} \tag{6}$$

Let $\mathbf{D}Q(\boldsymbol{\theta})$ and $\mathbf{D}^2Q(\boldsymbol{\theta})$ denote gradient and matrix of second partial derivatives of $Q(\boldsymbol{\theta})$, respectively. Finally let

$$\mathbf{U_n} = \frac{1}{n(n-1)} \sum_{i \neq j} g(\boldsymbol{\theta_0}, \mathbf{Z_i}, \mathbf{Z_j}),$$

The following theorem holds for a general convex function $f(\boldsymbol{\theta}, \mathbf{z_1}, \mathbf{z_2})$, not necessarily given by (5).

Theorem 2 *Assume that $Q(\boldsymbol{\theta})$ is twice differentiable at $\boldsymbol{\theta_0}$, moreover the matrix $\mathbf{H} = \mathbf{D}^2Q(\boldsymbol{\theta_0})$ is positive definite and there exists a neighborhood $\mathcal{B}$ of $\boldsymbol{\theta_0}$ such that $\mathbb{E}\,|g(\boldsymbol{\theta}, \mathbf{Z_1}, \mathbf{Z_2})|^2 < \infty$ for arbitrary $\boldsymbol{\theta} \in \mathcal{B}$. Then*

$$\sqrt{n}(\boldsymbol{\theta_n} - \boldsymbol{\theta_0}) \longrightarrow_d \mathcal{N}(\mathbf{0}, \mathbf{H}^{-1}\mathbf{G}\mathbf{H}^{-1}),$$

where

$$\mathbf{G} = \mathbb{E}\mathbf{A}\mathbf{A}^T, \quad \text{and} \quad \mathbf{A} = \mathbb{E}[g(\boldsymbol{\theta_0}, \mathbf{Z_1}, \mathbf{Z_2}) + g(\boldsymbol{\theta_0}, \mathbf{Z_2}, \mathbf{Z_1})|\mathbf{Z_1}].$$

Proof For simplicity we can assume that $\boldsymbol{\theta_0} = \mathbf{0}$ and $Q(\mathbf{0}) = 0$. Let us denote

$$T_n(\boldsymbol{\theta}, \mathbf{z_1}, \mathbf{z_2}) = f(\boldsymbol{\theta}/\sqrt{n}, \mathbf{z_1}, \mathbf{z_2}) - f(\mathbf{0}, \mathbf{z_1}, \mathbf{z_2}) - \frac{\boldsymbol{\theta}^T}{\sqrt{n}} g(\mathbf{0}, \mathbf{z_1}, \mathbf{z_2}).$$

It is easy to notice that at each point of differentiability of $Q(\boldsymbol{\theta})$ we have $\mathbf{D}Q(\boldsymbol{\theta}) = \mathbb{E}g(\boldsymbol{\theta}, \mathbf{Z_1}, \mathbf{Z_2})$, which implies $\mathbb{E}\, T_n(\boldsymbol{\theta}, \mathbf{Z_1}, \mathbf{Z_2}) = Q(\boldsymbol{\theta}/\sqrt{n})$. Moreover we define

$$V_n(\boldsymbol{\theta}) = \frac{1}{n(n-1)} \sum_{i \neq j} T_n(\boldsymbol{\theta}, \mathbf{Z_i}, \mathbf{Z_j}) = Q_n(\boldsymbol{\theta}/\sqrt{n}) - Q_n(\mathbf{0}) - \frac{\boldsymbol{\theta}^T}{\sqrt{n}} \mathbf{U_n},$$

which, for each fixed $\boldsymbol{\theta}$, is a U-statistic of order two with kernel T_n. Using inequality (6) and Lemma A, par. 5.2 in Serfling (1980), we can bound the variance of $V_n(\boldsymbol{\theta})$ from above as follows:

$$\begin{aligned} Var\ V_n(\boldsymbol{\theta}) &\leq \frac{2}{n}\, \mathbb{E}\, [T_n(\boldsymbol{\theta}, \mathbf{Z_1}, \mathbf{Z_2})]^2 \\ &\leq \frac{2}{n^2} \mathbb{E}\, [\boldsymbol{\theta}^T (g(\boldsymbol{\theta}/\sqrt{n}, \mathbf{Z_1}, \mathbf{Z_2}) - g(\mathbf{0}, \mathbf{Z_1}, \mathbf{Z_2}))]^2. \end{aligned}$$

Random variables $\boldsymbol{\theta}^T [g(\boldsymbol{\theta}/\sqrt{n}, \mathbf{Z_1}, \mathbf{Z_2}) - g(\mathbf{0}, \mathbf{Z_1}, \mathbf{Z_2})]$ are nonnegative and tend monotonically to random variable with expectation zero, because $\mathbf{D}Q(\boldsymbol{\theta})$ is continuous at zero. This implies that limiting variable is almost surely zero, which combined with the Lebesgue dominated convergence theorem shows that the variance of statistic $n\, V_n(\boldsymbol{\theta})$ tends to zero for every θ. Using Chebyshev inequality we get

$$n V_n(\boldsymbol{\theta}) - n Q(\boldsymbol{\theta}/\sqrt{n}) \to_P 0$$

for each $\boldsymbol{\theta}$. We can also use Taylor expansion to obtain

$$n Q(\boldsymbol{\theta}/\sqrt{n}) \to \frac{1}{2} \boldsymbol{\theta}^T \mathbf{H} \boldsymbol{\theta}.$$

Recapitulating we have

$$n Q_n(\boldsymbol{\theta}/\sqrt{n}) - n Q_n(\mathbf{0}) - \sqrt{n}\, \boldsymbol{\theta}^T \mathbf{U_n} - \frac{1}{2} \boldsymbol{\theta}^T \mathbf{H} \boldsymbol{\theta} \to_P 0.$$

Using Lemma 1 we get uniform convergence on compacts. For every $\varepsilon > 0$ and $M > 0$ the following inequality holds with probability at least $1 - \varepsilon$ for large n:

$$\sup_{|\boldsymbol{\theta}| \leq M} |n Q_n(\boldsymbol{\theta}/\sqrt{n}) - n Q_n(\mathbf{0}) - \sqrt{n}\, \boldsymbol{\theta}^T \mathbf{U_n} - \frac{1}{2} \boldsymbol{\theta}^T \mathbf{H} \boldsymbol{\theta}| < \varepsilon.$$

Random variable $-\sqrt{n}\, \mathbf{H}^{-1} \mathbf{U_n}$, which is the minimizer of the quadratic function $\sqrt{n}\, \boldsymbol{\theta}^T \mathbf{U_n} + \frac{1}{2} \boldsymbol{\theta}^T \mathbf{H} \boldsymbol{\theta}$, is bounded in probability. The same arguments as in Niemiro (1992, Theorem 4) imply that the limiting distributions of $\sqrt{n}\, \boldsymbol{\theta_n}$ and $-\sqrt{n}\, \mathbf{H}^{-1} \mathbf{U_n}$ are the same. Now it is enough to use CLT for U-statistics $\mathbf{U_n}$ (Serfling 1980, Theorem A, par. 5.5) to finish the proof. □

If the function f is again given by (5), then quick computation shows that its subgradient can be chosen as

$$g(\boldsymbol{\theta}, \mathbf{z_1}, \mathbf{z_2}) = (\mathbf{x_2} - \mathbf{x_1})\, \mathbb{I}(\boldsymbol{\theta}^T \mathbf{x_1} - \boldsymbol{\theta}^T \mathbf{x_2} < 1)\, \mathbb{I}(y_1 > y_2).$$

So the matrix $\mathbf{A}$ that was defined in Theorem 2 has the following form

$$\mathbf{A} = \mathbb{E}[(\mathbf{X_1} - \mathbf{X_2})\, \mathrm{sign}(\mathrm{Y}_2 - \mathrm{Y}_1)\, \mathbb{I}(\, \mathrm{sign}(\mathrm{Y}_2 - \mathrm{Y}_1)\, \boldsymbol{\theta_0}^{\mathrm{T}}(\mathbf{X_2} - \mathbf{X_1}) < 1) \mid \mathbf{Z_1}].$$

Moreover, let us assume (see Niemiro 1989, Theorem 2) that for every $\boldsymbol{\theta}$ in some neighborhood of $\boldsymbol{\theta_0}$ distributions of random variables $\boldsymbol{\theta}^T(\mathbf{X_1} - \mathbf{X_2})$ are absolutely continuous with densities $l(\boldsymbol{\theta}, t)$. Furthermore if conditional expectation

$$\mathbf{C}(\boldsymbol{\theta}, t) = \mathbb{E}\,[(\mathbf{X_1} - \mathbf{X_2})(\mathbf{X_1} - \mathbf{X_2})^T\, \mathbb{I}(Y_1 > Y_2) \mid \boldsymbol{\theta}^T(\mathbf{X_1} - \mathbf{X_2}) = t]$$

and l are continuous in some neighborhood of $(\boldsymbol{\theta_0}, 1)$, $l(\boldsymbol{\theta_0}, 1) > 0$ and matrix $\mathbf{C}(\boldsymbol{\theta_0}, 1)$ is positive definite then the matrix $\mathbf{H}$ from Theorem 2 equals

$$\mathbf{H} = l(\boldsymbol{\theta_0}, 1)\, \mathbf{C}(\boldsymbol{\theta_0}, 1).$$

References

Abrevaya J (1999) Computation of the maximum rank correlation estimator. Econ Lett 62:279–285

Bartlett PL, Ben-David S (2002) Hardness results for neural network approximation problems. Theor Comput Sci 284:53–66

Bloomfield P, Steiger WL (1983) Least absolute deviations: theory, applications, algorithms. Birkhäuser, Boston

Bobrowski L, Niemiro W (1984) A method of synthesis of linear discriminant function in the case of nonseparability. Pattern Recognit 17:205–210

Bobrowski L (2007) Ranked modelling of risk on the basis of survival data. ICSMRA, Lisbon

Bose A (1998) Bahadur representation of M_m estimates. Ann Stat 26:771–777

Clémençon S, Lugosi G, Vayatis N (2008) Ranking and empirical minimization of U-statistics. Ann Stat 36:844–874

Haberman SJ (1989) Concavity and estimation. Ann Stat 17:1631–1661

Han AK (1987) Non-parametric analysis of a generalized regression model. J Econ 35:303–316

Han AK (1988) Large sample properties of the maximum rank correlation estimator in generalized regression models. Preprint, Department of Economics, Harvard University, Cambridge, MA

Niemiro W (1989) L^1-optimal statistical discrimination procedures and their asymptotic properties. Mat Stos 31:57–89 (in Polish)

Niemiro W (1992) Asymptotics for M-estimators defined by convex minimization. Ann Stat 20:1514–1533

Niemiro W (1993) Least empirical risk procedures in statistical inference. Appl Math 22:55–67

Rockafellar RT (1970) Convex analysis. Princeton University Press, Princeton

Serfling RJ (1980) Approximation theorems of mathematical statistics. Wiley, New York

Sherman RP (1993) The limiting distributions of the maximum rank correlation estimator. Econometrica 61:123–137