

Improve NPL Analysis through
Pedigree-Selection: The Maximum
Selected Pedigree Subset-Score Method

Lars Ängquist¹

22nd February 2007

¹PhD Student at Centre for Mathematical Sciences, Department of Mathematical Statistics, Lund University, Lund, Sweden.

Abstract

In this article we develop and describe what we call the maximum selected pedigree subset-score method (MSPSSM). This method is intended to improve nonparametric linkage (NPL) analysis in the presence of disease heterogeneity, i.e. when only a subset of the pedigrees in the population actually is susceptible to the disease under study. Basically, one aims at reducing the noise to the linkage signal through forming test statistics using only a subset of the pedigrees in the pedigree set.

We perform simulations with respect to the MSPSS method, under various genetic parameter-settings and pedigree sets, and make appropriate comparisons to the standard NPL-score method. Moreover, we note on some similar alternative methods.

Key words: Nonparametric linkage analysis, maximum selected pedigree subset-score method, disease heterogeneity, significance levels and power, ROC-curves, selection criterias.

Contents

1	Introduction	3
1.1	The NPL Score	3
1.2	Outline of Paper	4
2	The MSPSS Method	4
2.1	Brief Motivation	5
2.2	General Formulation	5
2.3	Actual Test Statistic	6
3	Specific Formulations of Method	6
3.1	Threshold-Type Criterion	6
4	Results	7
4.1	Preliminaries	7
4.2	Significance Calculations Using the Threshold-Type Criterion	8
4.2.1	Score Distribution	8
4.2.2	Application to Affected Sib-Pairs	9
4.2.3	Main Analyses	11
4.2.4	Additional Analyses	11
5	Discussion	16
	References	16
A	Proportion-Type Criterion	18
B	Maximum Standardized Score Method	18
B.1	Basic Version	18
B.2	Revisited	19
B.3	Derivation of Criterion	20
C	Alternative Method	21

1 Introduction

The basis for linkage analysis is to observe marker data (MD), in the form of genotypes, with respect to a set of pedigrees, i.e. the so called *pedigree set*.¹ Procedures belonging to the field of *nonparametric linkage (NPL) analysis*, analyzing the inheritance of alleles with respect to the pedigree set and comparing what is actually found to what is expected under the null hypothesis H_0 of random inheritance, lead to a variety of statistical tests for the presence of *genetic linkage*. Note that such tests does not explicitly, but performance-wise implicitly, need assumptions regarding the underlying genetic disease model.²

Generally, for a single pedigree, the actual inheritance analysis is based on the concept of sharing alleles *identical-by-descent (IBD)*. To score inheritance patterns one introduces a *score function (S)* to numerically quantify the degree of sharing. The input to the score function is the so called *inheritance vector* (Donnelly, 1983; Kruglyak et al., 1996) rather than the marker data itself.³ As a general reference to NPL analysis consider, for instance, Ångquist (2007).

1.1 The NPL Score

Introduce the *pedigree-specific NPL score* at locus x as

$$Z(x) = \sum_{w \in \mathbb{V}} p_w(x) S(w), \quad (1)$$

where $p_w(x) = P[v(x) = w | \text{MD}]$ is the inheritance distribution at x , w is an inheritance vector, \mathbb{V} and MD are the set of all possible inheritance vectors and the marker data at x , and S is a standardized score function.⁴

Assume a pedigree set of N pedigrees. The (total or pedigree set) *NPL score* is generated as

$$Z(x) = \sum_{k=1}^N \gamma_k Z_k(x), \quad (2)$$

¹A pedigree is a, possibly multigenerational, connected set of relatives.

²This is why they are referred to as *nonparametric*!

³According to constructed (through the definition of S) or mandatory (inherent through v) symmetries, this basically corresponds to giving distinct scores to a certain number of defined different *IBD-sharing patterns* within the pedigree.

⁴Being *standardized* means that the original unstandardized scores (output from score function definition) have been transformed to obey $E(S|H_0) = 0$ and $V(S|H_0) = 1$.

where γ_k is the k^{th} pedigree weight, $\sum_{k=1}^N \gamma_k^2 = 1$ in order to preserve standardization (unit variance), and Z_k is the k^{th} pedigree-specific NPL score from (1).

For genome-wide scans, the obvious choice of *test statistic* is the maximum NPL score,

$$Z_{\max} = \max_x Z(x); \quad x \in \Omega, \quad (3)$$

where Ω is the genome region under study. Genome-wide approaches to significance calculations based on (3) is developed in e.g. Lander and Kruglyak (1995), Tang and Siegmund (2001) and Ängquist and Hössjer (2005) (analytical approximation), and Ott (1989), Malley et al. (2002) and Ängquist and Hössjer (2004) (Monte Carlo simulation).

For well-known alternative NPL-methods see e.g. Cordell et al. (1995) and Kong and Cox (1997). A brief overview of several test statistics is given in Haines and Pericak-Vance (1998).

1.2 Outline of Paper

In *Section 2* we present the motivation for a population-based disease heterogeneity linkage score and formulate the general formulation of the method based on the nonparametric MSPSS-score, which is a function of the ordered set of pedigree-specific NPL scores and their corresponding weights. Next, *Section 3* is dedicated to the method-instance based on a threshold-type selection criterion, i.e. which is defined through selecting pedigree-specific scores with respect to a predefined score criterion. Further, in *Section 4* we perform Monte Carlo simulations in order to calculate significance levels and power, displaying results through ROC-curves, and some final discussion is given in *Section 5*. Some alternative (perhaps peripheral) methods, and extensions, are referred to the *Appendices*

2 The MSPSS Method

Here we will describe the new algorithmic procedure called the *maximum selected pedigree subset-score method (MSPSSM)*.

2.1 Brief Motivation

Standard linkage analysis-procedures are based on the assumption of, or are at least designed for, observing a pedigree set from a *homogeneous* population, i.e. where the underlying genetic model is a valid disease description for *all* observable pedigrees constituting the population of interest. One might argue that exceptions to such a truth may exist. Explicitly, these deviations may correspond to: (i) Locus heterogeneity, i.e. that different influential disease loci for different subpopulations are present. This might be of severe importance regarding the context of complex diseases. (ii) That the disease exists only in a specific subpopulation.

The MSPSS method outlined below is a nonparametric approach designed to being well-adapted to such cases, as a way of partially dealing with this complicating fact. For information, and further references, on parametric approaches to linkage analysis incorporating similar heterogeneity-parameters consider e.g. Ott (1999).

2.2 General Formulation

Assume a pedigree set consisting of N pedigrees. Let Z_1, Z_2, \dots, Z_N be the N pedigree-specific NPL scores (1) weighted through (2). Permute this sequence to form the ordered scores $Z_{(1)}, Z_{(2)}, \dots, Z_{(N)}$, i.e. where $Z_{(k)}$ is the k^{th} largest score and, consequently, $Z_{(1)} \geq Z_{(2)} \geq \dots \geq Z_{(N)}$.

Now, pick the n largest pedigree-specific scores from (2) and calculate the corresponding n^{th} *pedigree subset-score (PSS)*

$$Z^n(x) = \sum_{k=1}^n \gamma_{(k)} Z_{(k)}(x); \quad 1 \leq n \leq N. \quad (4)$$

Let us now generally define the *selected pedigree subset-score (SPSS)* as

$$Z^{n'}(x) = \sum_{k=1}^{n'} \gamma_{(k)} Z_{(k)}(x); \quad 1 \leq n' \leq N, \quad (5)$$

where (in the most general form) $n' = g(\{Z_k\}, \{\gamma_k\})$ is chosen with respect to a selection criterion function $g(\cdot)$ based on *both* all the pedigree-specific scores and the corresponding pedigree-weights.

2.3 Actual Test Statistic

The intuitive and most natural test statistic using the formulation of (5), with respect to a genome-wide study over region Ω , is

$$Z_{\max}^{n'} = \max_x Z^{n'}(x); \quad x \in \Omega, \quad (6)$$

i.e. the *maximum selected pedigree subset-score (MSPSS)*.

According to problems with computational complexity it might seem favourable, in some cases, to approximate (6) with calculating (5) only with respect to either (i) loci $\{x; Z(x) \geq T\}$ or (ii) loci $\{x; x = \arg Z_{\max}\}$.

3 Specific Formulations of Method

We mainly consider a score threshold-based instance of (5) which is independent of the pedigree γ -weights. Note that it is obviously not possible, using (5), to directly numerically compare results based on distinct thresholds. Hence we are forced to simulate and calculate corresponding significance calculations *for each case separately*. Using so called *receiver operating characteristics (ROC)-curves* (Selin, 1965; Bradley, 1996) facilitates simultaneous plotting, power versus significance levels, and hence comparisons between cases.

Remark 1 *As a side effect, such formulations of (5) implies that it is no longer important to standardize the γ -weights above, i.e. the set of weights only express relative importance between the pedigrees in the pedigree set, without any corresponding forced numerical scale-restriction.*

3.1 Threshold-Type Criterion

Choose n' as the number of pedigree-specific scores *larger than threshold T* , i.e. $n' = |\{Z_k; Z_k \geq T\}|$. For pedigree k , the threshold may be explicitly predefined or implicitly through a given, exact or approximate, quantile of the score null distribution $F_{H_0}^k$ (based on score function S_k).

Remark 2 *Let us assume knowledge of the underlying genetic model, i.e. complete understanding of the valid (present) alternative hypothesis H_1 , which*

in turn facilitates computation of distribution $\bar{F}_{H_1} = 1 - F_{H_1}$. In this case, reasonable choices of T may be based on, for example,

$$\begin{cases} T = \arg \max_x [\bar{F}_{H_1}(x) - \bar{F}_{H_0}(x)], \\ T = \arg \max_x ([\bar{F}_{H_1}(x) - \bar{F}_{H_0}(x)] / F_{H_0}(x)). \end{cases}$$

Remark 3 As the threshold selection-rules in Remark 2 suggests it is perfectly possible to use different thresholds for different pedigrees in the same pedigree set. Assuming perfect data, our suggestion is to reserve this alternative to different pedigree units in a nonhomogeneous pedigree set.⁵

4 Results

In this section we will perform and discuss one-locus significance calculations, significance levels (α) and power (β), with respect to the *homogeneous* pedigree-sets based on units displayed in Figure 1.⁶ Pedigrees 1-3 is artificial, but reasonably interesting and instructive, pedigree units, and Pedigrees 4-5 are real example units taken from the BOTNIA-study (Lindgren et al., 2002; Ängquist and Hössjer, 2005).

4.1 Preliminaries

Throughout we consider a genome based on a single chromosome $\Omega = C$, with length $|C| = 3$ Morgans, and equal weighting (all γ -weights equal) with respect to pedigree sets based on $N = 50$ or $N = 200$ pedigrees

We define alternative hypotheses through the following *genetic disease models*: (i) Assuming disease allele frequency $p = p(D) = 0.01$. (ii) Using any of the penetrance vectors,

$$\begin{cases} f^1 = (0.001, 0.999, 0.999), \\ f^2 = (0.001, 0.500, 0.999), \\ f^3 = (0.001, 0.001, 0.999), \end{cases} \quad (7)$$

where $f = (f_0, f_1, f_2)$ describes the connection between the phenotype (disease status) and genotype (disease- and nondisease alleles) at the disease

⁵See Footnote 6 below.

⁶Explicitly, in a *homogeneous* pedigree set all pedigrees in the set has common structure and phenotype setting (given by the unit).

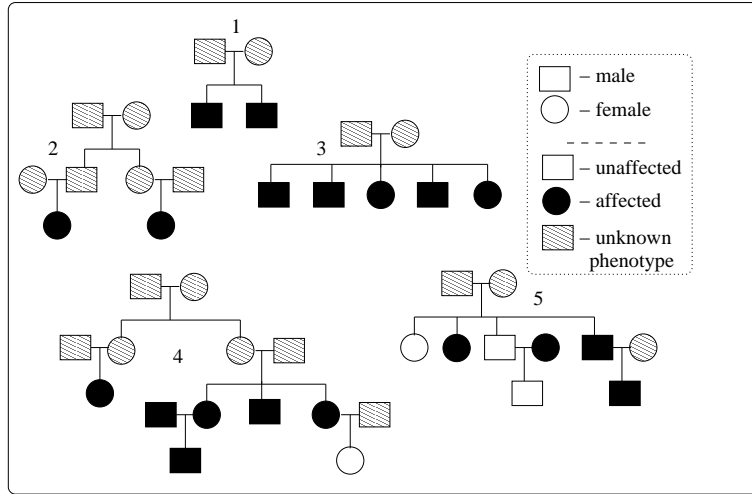


Figure 1: A pedigree set consisting of 5 different pedigrees, all with distinct structures and phenotype settings. Note that these pedigrees function as our basic pedigree units, forming the basis for our homogeneous pedigree sets used in the analyses.

locus. Explicitly, f_i equals the probability of being affected given a disease genotype consisting of i D s and $2-i$ d s. (iii) Assuming that the disease locus l is located in the middle of the first (and only) chromosome ($l = 1.5M$).

Further, we assume *perfect marker data*, i.e. fully observable inheritance vectors through the complete genome, base our analyses on pedigree-specific scores produced by the score function S_{pairs} (Whittemore and Halpern, 1994) and compare the NPL score-method versus the threshold-type MSPSS-method through plotting ROC-curves.

Remark 4 Note: (i) All significance calculations are performed using crude Monte Carlo simulations. (ii) We adopt (6), i.e. we calculate the MSPSS-statistic at all loci throughout Ω .

4.2 Significance Calculations Using the Threshold-Type Criterion

4.2.1 Score Distribution

For a pedigree unit the score distribution is based on all possible pedigree-specific scores, where each score corresponds to a specific inheritance vector

v .⁷ Note that the distribution clearly depends on choice of score function, the pedigree structure and present phenotype setting. Facing imperfect data makes the distribution dependent on marker data from the founders, which otherwise is not the case.

In our main analyses we tune the selection criteria with score threshold $T = 1.25$. Basically this means that, in the MSPSS case, we base each analysis only on pedigree scores which exceed threshold 1.25. For more information on the relation between this specific threshold and the five marginal (pedigree unit) distributions, observing $\bar{F}_{H_0}(1.25)$, for the pedigrees in Figure 1, see Table 1 and Figure 2.⁸

Table 1: Probabilities, for Pedigrees 1-5, $\bar{F}_{H_0}(1.25)$ of marginally exceeding the threshold $T = 1.25$ under the null hypothesis of no linkage. [To ease interpretation of $\bar{F}_{H_1}(1.25)$ under the alternative hypothesis, see this as the marginal distribution at the (under the assumptions false) disease locus.] Moreover the number of possible inheritance vectors $|\mathbb{V}|$ with respect to the set of all inheritance vectors \mathbb{V} is shown.

Pedigree	$ \mathbb{V} $	$\bar{F}_{H_0}(1.25)$
1	$2^4 = 16$	0.25
2	$2^8 = 256$	0.25
3	$2^{10} = 1024$	0.1211
4	$2^{16} = 65536$	0.1416
5	$2^{12} = 4096$	0.25

4.2.2 Application to Affected Sib-Pairs

The pedigree unit corresponding to Pedigree 1 in Figure 1 is called an *affected sib-pair (ASP)* and is a common unit with respect to genetic analysis. Here the number of inheritance vectors $|\mathbb{V}|$ equals 16, but the number of distinct

⁷In other words, the distribution is based on the set of scores $\{S\} = \{S(w); w \in \mathbb{V}\}$, where \mathbb{V} is the set of possible inheritance vectors.

⁸Note that all subgraphs are based on the *founder couple reduction*-reduced set of inheritance vectors (Kruglyak et al., 1996; Gudbjartsson et al., 2000), i.e. $|\mathbb{V}| = 2^{m-f}$ rather than 2^m , where m and f are the number of meioses and founders.

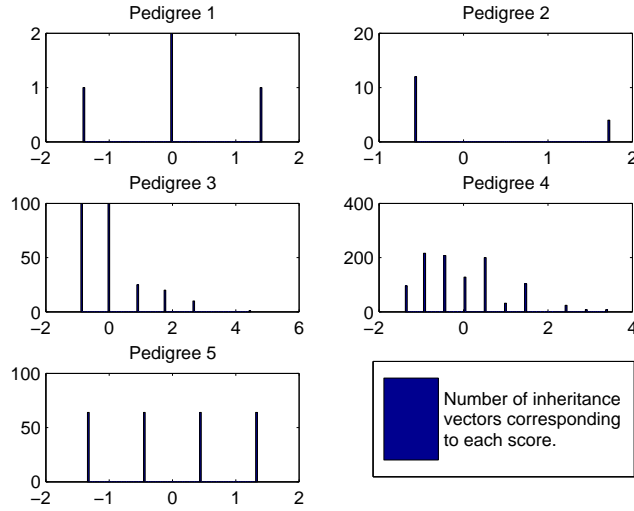


Figure 2: The pedigree-specific marginal score distributions for Pedigrees 1-5. For more information, see Table 1.

scores is only three. Using S_{pairs} one finds the standardized score vector $(s_0, s_1, s_2) = (-\sqrt{2}, 0, \sqrt{2})$ corresponding to the ASP sharing 0, 1 or 2 alleles IBD respectively.⁹

This implies that choosing $0 < T \leq \sqrt{2} \approx 1.4142$ virtually equals using a test statistic based on the number of ASPs sharing 2 alleles IBD.

Remark 5 *Note that since the test is degenerate for $T > \sqrt{2}$ and $\bar{F}_{H_0}(T) = 0.25$ for $0 < T \leq \sqrt{2}$, which is quite large, we will in an expected value-sense always pick at least 25% of the nonsusceptible pedigree-specific scores when calculating (5). Sadly, this means that it is impossible to further reduce the present noise (noninformative variance) from the linkage signal (with respect to the test statistic).*

⁹The same is true, for instance, for the well-known score function S_{all} . In fact, all score functions *equivalent* to S_{pairs} and S_{all} produces the same score distribution for ASPs. Equivalence, assuming nonstandardized scores with $s'_2 \geq s'_1 \geq s'_0$, may be formalized through the equivalence class fulfilling $(s'_2 - s'_1) = c(s'_1 - s'_0)$ for a given constant $c \geq 0$. The class including S_{pairs} and S_{all} corresponds to all the *symmetric* score functions where $c = 1$. For some discussion on score functions for ASPs consider e.g. Ängquist et al. (2005).

4.2.3 Main Analyses

Let us introduce the heterogeneity parameter h which equals the probability for a randomly chosen (ascertained) pedigree *to be susceptible* for the disease, i.e. that the genetic model corresponding to an alternative hypothesis is valid for this specific pedigree.¹⁰

Remark 6 *Note that $h = 1$ correspond to a homogeneous population, whereas $h = 0$ relates to a nonexisting disease (in this population) and $0 < h < 1$ describes the strength of the disease heterogeneity (with respect to this population); increasing strength for decreasing h .*

In the main analyses we set $N = 50$ and $h = 0.2$. The results, for each penetrance setting separately, are displayed in Figures 3-5. The MSPSS-method generally performs best, and is favourable to the NPL-method, for Pedigrees 3-4, i.e. in these cases where the number of distinct scores is the largest and probability $\bar{F}_{H_0}(T)$ is the smallest. In the other cases the performance is similar or, sometimes, even in favour of the NPL-approach.

4.2.4 Additional Analyses

In the additional analyses we set $N = 200$ and use penetrance vector f^1 for Pedigree 4. We set up diverse settings by combining thresholds $T \in \{1.25, 2.5\}$ and heterogeneity $h \in \{1, 0.2, 0.1, 0.05\}$ in all possible ways. For the relation between the two thresholds and the corresponding null- and alternative marginal distributions (and their ratios), see Table 2.

Table 2: Probabilities, for Pedigree 4, for marginally exceeding $T = 1.25$ and $T = 2.5$ under null- and alternative hypothesis based on penetrance f^1 .

T	$\bar{F}_{H_0}(T)$	$\bar{F}_{H_1}(T)$	$\bar{F}_{H_1}(T)/\bar{F}_{H_0}(T)$
1.25	0.1416	0.7861	5.5518
2.5	0.0156	0.2048	13.1123

The results are displayed in Figure 6.

¹⁰In the Monte Carlo simulations under H_1 , we first define N and then randomly generate the number of susceptible pedigrees N_h according to generating a random number from a binomial distribution, i.e. $N_h \sim Bin(N, h)$.

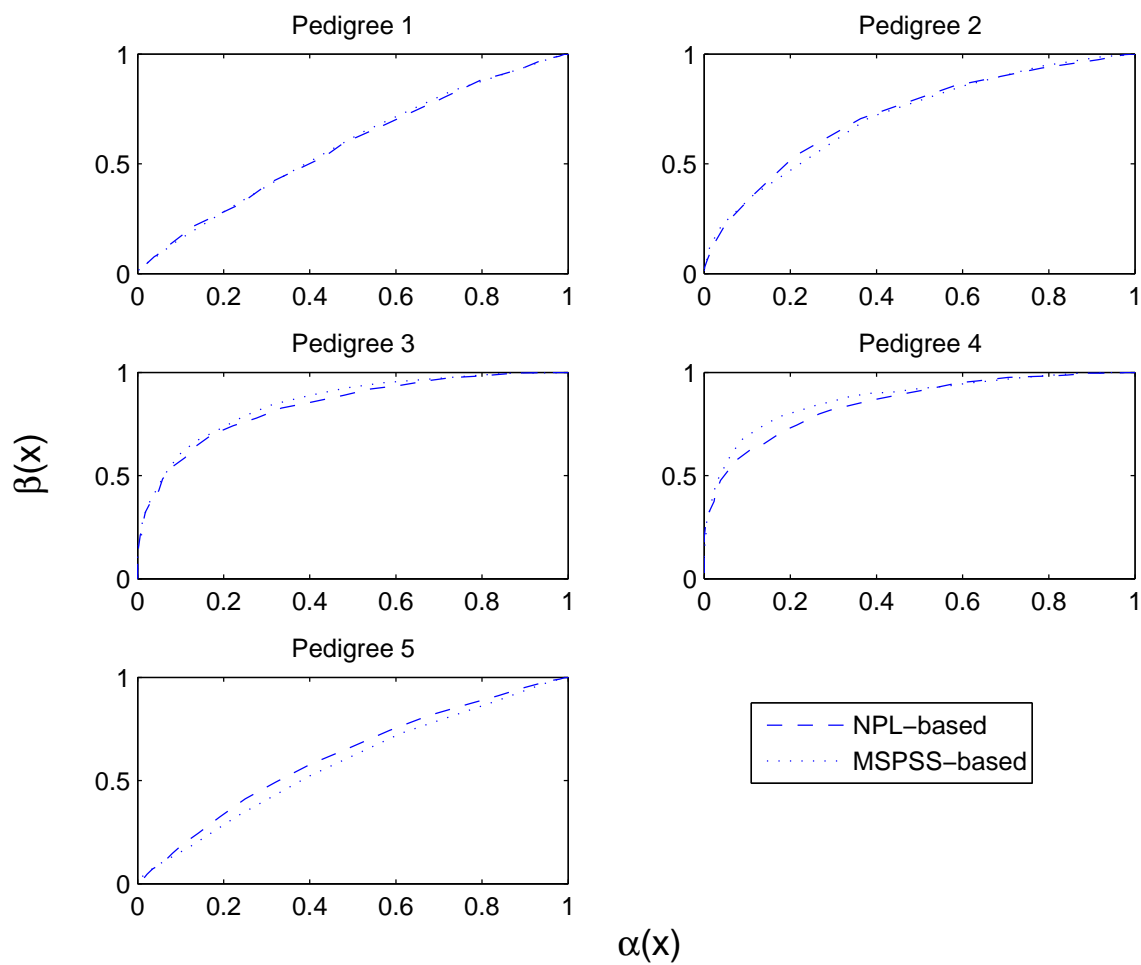


Figure 3: Comparing the NPL-based and the MSPSS-based methods, under penetrance setting f^1 , using threshold $T = 1.25$, number of pedigrees $N = 50$, heterogeneity $h = 0.2$ and number of simulations $J = 1000$.

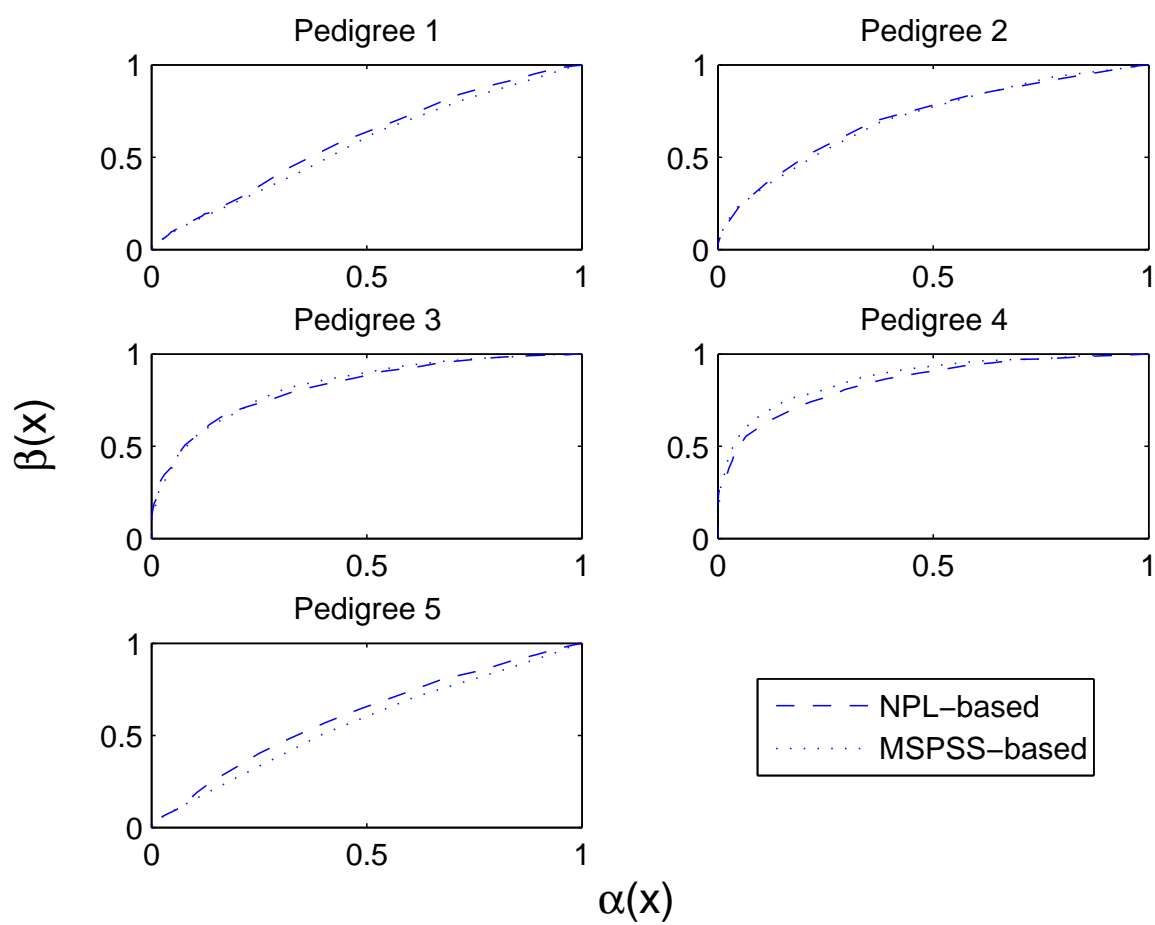


Figure 4: Comparing the NPL-based and the MSPSS-based methods, under penetrance setting f^2 . For more information, see Figure 3.

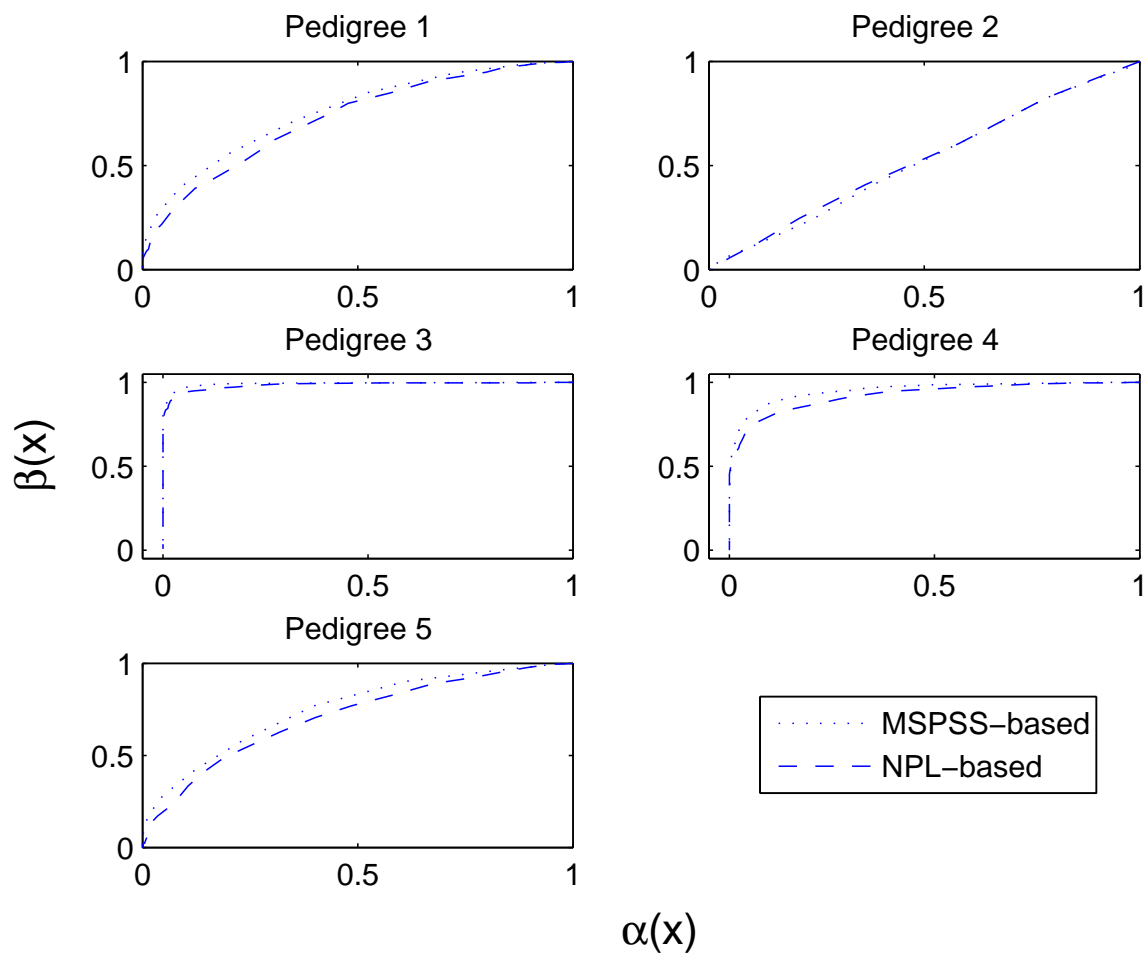


Figure 5: Comparing the NPL-based and the MSPSS-based methods, under penetrance setting f^3 . For more information, see Figure 3.

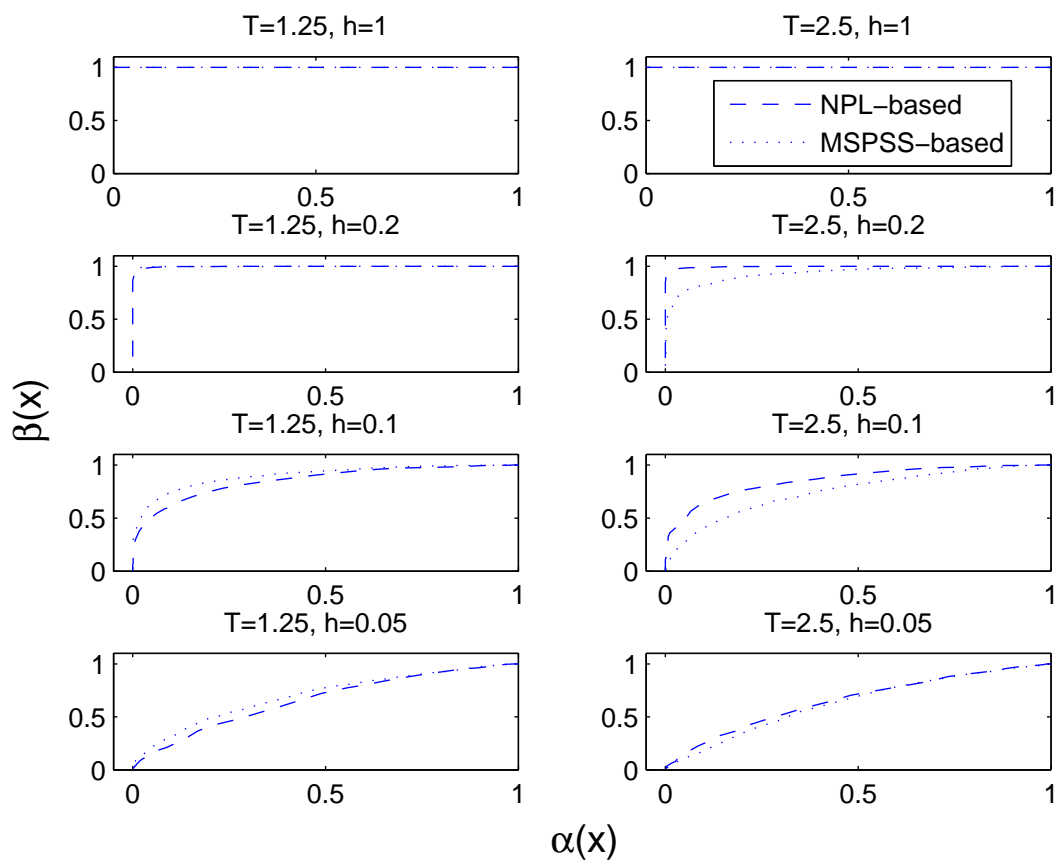


Figure 6: Comparing the NPL-based and the MSPSS-based methods, under penetrance setting f^1 , using thresholds $T \in \{1.25, 2.5\}$, number of pedigrees $N = 200$, heterogeneity parameters $h \in \{0.05, 0.1, 0.2, 1\}$ and number of simulations $J = 1000$.

5 Discussion

Looking at the Figure 6 one might note that generally is in favour of the MSPSS-method for $T = 1.25$, whereas the opposite conclusion is true for the larger threshold $T = 2.5$. With decreasing h both methods loose power, which is fully consistent with the underlying theory (and intuitive understanding). The problem, for the MSPSS-method, when $T = 2.5$ is that the number of selected susceptible pedigrees probably is very low (though the ratio between the number of susceptible and nonsusceptible pedigrees should be high) which leads to loss of power. Using the NPL-method keeps all the susceptible pedigrees but, at the same time, the linkage signal is blurred according to the nonsusceptible counterparts.

To sum up, properly tuned (using T with respect to given N and assumed h) the MSPSS-method might in many cases be more powerful than the conventionally used NPL-score.

References

- Ängquist, L. (2007). *Pointwise and genomewide significance calculations in gene mapping through nonparametric linkage analysis: Theory, algorithms and applications* (Doctoral thesis No. 2006:15). Lund: Department of Mathematical Statistics, Lund University.
- Ängquist, L., Anevski, D. and Luthman, H. (2005). *Unconditional two-locus nonparametric linkage analysis: On composite null hypotheses with and without gene-gene interaction* (Tech. Rep. No. 2005:28). Lund: Department of Mathematical Statistics, Lund University.
- Ängquist, L. and Hössjer, O. (2004). Using importance sampling to improve simulation in linkage analysis. *Statistical Applications in Genetics and Molecular Biology*, 3(1:5). (Electronic journal, 24 pages)
- Ängquist, L. and Hössjer, O. (2005). Improving the calculation of statistical significance in genome-wide scans. *Biostatistics*, 6(4), 520–538.
- Bradley, A. P. (1996). ROC curves and the χ^2 test. *Pattern Recognition Letters*, 17, 287–294.

- Cordell, H. J., Todd, J. A., Bennett, S. T., Kawaguchi, Y. and Farrall, M. (1995). Two-locus maximum lod score analysis of a multifactorial trait: Joint consideration of IDDM2 and IDDM4 with IDDM1 in type 1 diabetes. *American Journal of Human Genetics*, *57*, 920–934.
- Donnelly, K. P. (1983). The probability that related individuals share some section of the genome identical by descent. *Theoretical Population Biology*, *23*, 34–64.
- Gudbjartsson, D. F., Jonasson, K., Frigge, M. and Kong, A. (2000). ALLEGRO, a new computer program for multipoint linkage analysis. *Nature Genetics*, *25*, 12–13.
- Haines, J. L. and Pericak-Vance, M. A. (1998). *Approaches to gene mapping in complex human diseases*. New York: John Wiley & Sons.
- Kong, A. and Cox, N. (1997). Allele-sharing models: LOD scores and accurate linkage tests. *American Journal of Human Genetics*, *61*, 1179–1188.
- Kruglyak, L., Daly, M. J., Reeve-Daly, M. P. and Lander, E. S. (1996). Parametric and nonparametric linkage analysis: A unified multipoint approach. *American Journal of Human Genetics*, *58*, 1347–1363.
- Lander, E. S. and Kruglyak, L. (1995). Genetic dissection of complex traits: Guidelines for interpreting and reporting linkage results. *Nature Genetics*, *11*, 241–247.
- Lindgren, C. M., Mahtani, M. M., Widén, E., McCarthy, M. I., Daly, M. J., Kirby, A., Reeve, M. P., Kruglyak, L., Parker, A., Meyer, J., Almgren, P., Lehto, M., Kanninen, T., Tuomi, T., Groop, L. C. and Lander, E. S. (2002). Genomewide search for type 2 diabetes mellitus susceptibility loci in Finnish families: The BOTNIA study. *American Journal of Human Genetics*, *70*, 509–516.
- Malley, J. D., Naiman, D. and Bailey-Wilson, J. (2002). A comprehensive method for genome scans. *Human Heredity*, *54*, 174–185.
- Ott, J. (1989). Computer-simulation methods in human linkage analysis. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(11), 4175–4178.

Ott, J. (1999). *Analysis of human genetic linkage* (Third ed.). New York: The John Hopkins University Press.

Selin, I. (1965). *Detection theory*. Princeton, New Jersey: Princeton University Press.

Tang, H. K. and Siegmund, D. (2001). Mapping quantitative trait loci in oligogenic models. *Biostatistics*, 2, 147–162.

Whittemore, A. S. and Halpern, J. (1994). A class of tests for linkage using affected pedigree members. *Biometrics*, 50, 118–127.

A Proportion-Type Criterion

Instead of selecting pedigrees using a criterion based on predefined score thresholds one may directly define a fixed (nonrandom) number n' .¹¹ Here the same kind of judgements as in the threshold-case must be made, i.e. one must consider the pedigree null hypothesis and an assumption on the alternative hypothesis and make a decision. Of course, this approach has some (rather vague) connection to the assumed proportion of susceptible pedigrees in the underlying population, but interpretational power is diminished since in this case we may not even easily calculate the expected number of selected nonsusceptible pedigrees under given H_1 .¹²

Remark 7 *One may note that using fixed n' (proportion-based selection) makes the selection procedure independent not only of pedigree-weights but even of the pedigree-specific scores.*

B Maximum Standardized Score Method

B.1 Basic Version

Let us take the n largest pedigree-specific scores and their corresponding γ -weights from (2) and define M_n as

$$M_n(x) = \sum_{k=1}^n \gamma_{(k)}^n Z_{(k)}(x); \quad 1 \leq n \leq N, \quad (8)$$

¹¹For example, computing the sum in (5) over the top 0.1%, 1%, 5% or 10% scores.

¹²Calculations involve messy derivations through using *order statistics*.

where the adjusted weights $\sum_{k=1}^n [\gamma_{(k)}^n]^2 = \sum_{k=1}^n \gamma_{(k)}^2 / C_n = 1$ and C_n is a normalizing constant. Using (8) makes it possible to compute the *maximum standardized score (MSS)*

$$M(x) = \max_n M_n(x) = \max_n \sum_{k=1}^n \gamma_{(k)}^n Z_{(k)}(x); \quad 1 \leq n \leq N. \quad (9)$$

When defining a test statistic it is more natural to use the $\max_{x \in \Omega} M(x)$ based on the actual (reweighted) MSS score (9) than the corresponding version of (6).¹³

Remark 8 *This method may be motivated by the fact that one chooses the pedigree scores (and the locus) that, pretending that the complementary information never existed, produces the maximum traditional NPL-score. However, sadly but true, this seems more to be a fun note than a very strong argument.*

B.2 Revisited

Reformulating the criterion above we assume an underlying locus and oppress the dependence on x throughout this argument. We start out using (8),

$$\left\{ \begin{array}{l} M_n = \sum_{k=1}^n \gamma_{(k)}^n Z_{(k)}, \\ M_{n+1} = \sum_{k=1}^n \gamma_{(k)}^{n+1} Z_{(k)}(x) + \gamma_{(n+1)}^{n+1} Z_{(n+1)}, \\ M_{n+1} - M_n = \sum_{k=1}^n [\gamma_{(k)}^{n+1} - \gamma_{(k)}^n] Z_{(k)}(x) + \gamma_{(n+1)}^{n+1} Z_{(n+1)}, \end{array} \right. \quad (10)$$

¹³The former case may be defined through $Z_{\max}^{n'}$ where $n' = \arg M = \arg \max_n M_n$. The argument is supported by n' not necessarily increasing with respect to, for instance, strength of genetic model under H_1 and proportion of susceptible pedigrees in population; calling for reweighting of pedigrees.

Using the third part of (10) one may note that

$$\begin{aligned}
M_{n+1} - M_n \geq 0 &\Leftrightarrow Z_{(n+1)} \geq \frac{1}{\gamma_{(n+1)}^{n+1}} \sum_{k=1}^n \left[\gamma_{(k)}^n - \gamma_{(k)}^{n+1} \right] Z_{(k)}(x) \\
&\Leftrightarrow Z_{(n+1)} \geq \frac{1}{\gamma_{(n+1)}^n} \left[\sqrt{1 + \left[\gamma_{(n+1)}^n \right]^2} - 1 \right] \sum_{k=1}^n \gamma_{(k)}^n Z_{(k)}(x) \\
&\Leftrightarrow h(n+1)M_n,
\end{aligned} \tag{11}$$

where the function $h(n+1) = h(n+1; \{Z_k\}, \{\gamma_k\})$ implicitly depends on the original pedigree-specific weights and scores from (2). For derivation of the final equivalence in (11), see Section B.3.

In other words the $(n+1)^{\text{th}}$ score (8) is favourable with respect to the n^{th} score if (11) is satisfied.

Remark 9 *Since the right-most side (sum) in (11) is not guaranteed to increase it is not possible to exclude multiple local maximums of the M_n -function.*

B.3 Derivation of Criterion

Let $\gamma_{(n+1)}^n$ be the $(n+1)^{\text{th}}$ pedigree-weight (normalized with C_n). Further,

$$\sum_{k=1}^{n+1} \left[\gamma_{(k)}^n \right]^2 = \sum_{k=1}^n \left[\gamma_{(k)}^n \right]^2 + \left[\gamma_{(n+1)}^n \right]^2 = 1 + \left[\gamma_{(n+1)}^n \right]^2,$$

which implies, that in order to preserve the standardization, we need to redefine the weights as

$$\gamma_{(k)}^{n+1} = \gamma_{(k)}^n / \sqrt{1 + \left[\gamma_{(n+1)}^n \right]^2}; \quad k = 1, 2, \dots, n+1. \tag{12}$$

Now, inserting (12) we have

$$\begin{aligned}
\frac{1}{\gamma_{(n+1)}^{n+1}} \left[\gamma_{(k)}^n - \gamma_{(k)}^{n+1} \right] &= \left[\frac{\sqrt{1 + \left[\gamma_{(n+1)}^n \right]^2}}{\gamma_{(n+1)}^n} \right] \left[\gamma_{(k)}^n - \frac{\gamma_{(k)}^n}{\sqrt{1 + \left[\gamma_{(n+1)}^n \right]^2}} \right] \\
&= \frac{1}{\gamma_{(n+1)}^n} \left[\sqrt{1 + \left[\gamma_{(n+1)}^n \right]^2} - 1 \right] \gamma_{(k)}^n,
\end{aligned}$$

where the first two terms in the final product are independent of k .

C Alternative Method

As an alternative to the selection procedures outlined in Section 3 and Appendix A, it is possible to focus on *pedigree-proportion of total score* $Z(x)$ in (2) rather than pedigree-specific scores in (1), i.e. to define the ordered score sequence $Z_{(1')}, Z_{(2')}, \dots, Z_{(N')}$ through the alternative ordering-relation

$$\gamma_{(1')}Z_{(1')} \geq \gamma_{(2')}Z_{(2')} \geq \dots \geq \gamma_{(N')}Z_{(N')}.$$

Now, the alternative procedures are readily available by replacing scores $Z_{(k)}$ with the products $Z_{(k')}$.¹⁴ This approach may be interpreted as incorporating *a priori information* on relative importance of pedigrees into the analysis.

¹⁴Note that $[(1'), (2'), \dots, (N')]$ is a permutation of $[(1), (2), \dots, (N)]$, which in itself is yet another permutation of the standard increasing counting sequence $[1, 2, \dots, N]$.