View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Göteborgs universitets publikationer - e-publicering och e-arkiv

CHALMERS

GÖTEBORGS UNIVERSITET

Storskalig nätverksestimering

Utvärdering av en ny metod för glesa nätverk

Examensarbete för kandidatexamen i Matematik vid Göteborgs universitet

Andersson, Jenny
Bertilsson, Rebecka
Foogde, Helena
Köllerström, Lovisa
Lindström, Robin

Institutionen för Matematiska Vetenskaper
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2018

Storskalig nätverksestimering

Utvärdering av en ny metod för glesa nätverk

Examensarbete för kandidatexamen i Matematisk statistik vid Göteborgs universitet

Jenny Andersson Rebecka Bertilsson Helena Foogde Robin Lindström

Examensarbete för kandidatexamen i Matematisk statistik inom Matematikprogrammet vid Göteborgs universitet

Lovisa Köllerström

Handledare: Rebecka Jörnsten
Examinator: Marina Axelsson-Fisk
Maria Roginskaya

Institutionen för Matematiska Vetenskaper
GÖTEBORGS UNIVERSITET
Göteborg, Sverige 2018

Förord

Författarna till denna rapport har under arbetets gång fört loggbok över den tid som respektive författare har lagt på att arbeta med detta kandidatarbete. Gemensamt har även en dagbok skrivits för att dokumentera och planera projektet.

Bidragsrapport
I detta kandidatarbete har alla författare deltagit i varje del av processen. Inom gruppen har vi ansett det viktigt att alla har lärt sig, förstått och hjälpt till i varje arbetssteg, från inhämtning av fakta till programmering och analys av resultat. Gruppen har arbetat tätt ihop och har därmed kunnat planera arbetet och diskutera problem kontinuerligt. Helena Foogde och Robin Lindström har haft huvudansvar för kodning och simulering, vilket innebär att de har lagt ner mer tid än övriga gruppmedlemmar i denna del av arbetet. På liknande sätt har Jenny Andersson, Rebecka Bertilsson och Lovisa Köllerström haft huvudansvar för skrivandet av rapporten. Uppdelningen har fallit sig naturlig under arbetets gång på grund av intresse.

Tack
Vi vill ge ett stort tack till vår handledare Rebecka Jörnsten för all hjälp och vägledning som vi fått under arbetets gång. Vi vill också tacka Anna Källsgård för hennes stöd under rapportskapandet och synpunkter på presentationen.

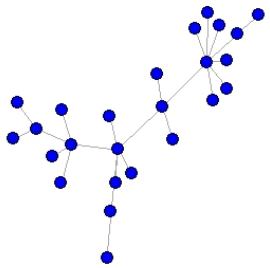
Populärvetenskaplig presentation

Utvärdering av en ny metod för tillämpning inom exempelvis cancerforskning

Detta projekt utvärderar en ny metod vid namn *k-glasso*, som används för uppskattning av glesa nätverk av stora dimensioner. Slutsatsen från arbetet är att metodens framgång är beroende av hur datan som används för att uppskatta nätverket ser ut.

Inom många forskningsområden förekommer komplexa system. En nätverksmodell är ett förenklat sätt att beskriva sådana system, se Figur 1. Vi kan visualisera ett nätverk på följande sätt: föreställ dig olika vägar mellan städer, där städerna är de objekt som studeras och de kopplingar som finns mellan objekten är vägar som sammanbinder städerna. Oftast är det ointressant huruvida det går att ta sig från en stad till en annan via andra städer. Det intressanta är istället om det går att ta sig direkt mellan två städer utan att åka via en tredje stad.

Ett annat exempel, från cancerforskningen, är hur olika gener samspelar med varandra. Målet med skattning av ett sådant nätverk är då att hitta de gener som påverkar varandra direkt. I förlängningen kan dessa skattningar leda till större förståelse för olika typer av cancer och i bästa fall leda till framsteg inom forskningen.



Figur 1: *Ett nätverk.*

I verkligheten är det svårt att ta reda på exakt hur ett nätverk ser ut. Metoder för att uppskatta olika typer av 'enklare' nätverk är välfungerande, men inom exempelvis cancerforskning har nätverken ofta stora dimensioner och är glesa, vilket leder till svårare nätverksskattningar. Tänk dig att du har information om 20 olika gener från 100 patienter, i jämförelse med om du har information om 50 000 gener från 100 patienter. Det sistnämnda fallet är ett komplext problem, eftersom antalet gener är fler än antalet patienter.

I artikeln "*Improving the Graphical Lasso Estimation for the Precision Matrix Through Roots of the Sample Covariance Matrix*" [1] som publicerades hösten 2017, presenterades ett nytt tillvägagångssätt för att hantera liknande problem. Artikelns grundidé är att ta kvadratroten ur datan innan uppskattningsmetoden appliceras. Deras slutsatser var att denna transformation dels bidrar till en stabilare skattning av det sanna nätverket än tidigare metoder, och dels till att beräkningstiden reduceras kraftigt. Resultaten från vårt projekt stöttar inte dessa slutsatser fullt ut, då metoden verkar vara databeroende.

Sammanfattning

Partiell korrelation mellan variabler kan implicit erhållas genom precisionsmatrisen. För att estimera denna kan den empiriska kovariansmatrisen inverteras. Problem uppstår när antalet variabler p är större än antalet observationer n, eftersom kovariansmatrisen då får låg rang och inte kan inverteras. Tidigare metoder för att lösa detta problem är osäkra och därför har en metod vid namn *k-glasso* utvecklats. I oktober 2017 publicerades artikeln "*Improving the Graphical Lasso Estimation for the Precision Matrix Through Roots of the Sample Covariance Matrix*" [1], där k-glasso presenterades. I artikeln konstaterades att denna metod presterar bättre än föregående metoder. Syftet med denna studie var att undersöka hur väl k-glasso presterar i att estimera stora glesa precisionsmatriser som liknar nätverk från tillämpningar. I simuleringen genererades två blockdiagonala precisionsmatriser för olika värden på p, där de underliggande nätverken hade en fördelning av typ *scale-free*. Dessutom undersöktes en nätverksmodell från originalartikeln i två variationer. Den k:te roten ur den empiriska kovariansmatrisen beräknades genom att ta k-roten ur diagonalen i dess egenvärdesdekomposition. R-funktionen *huge()* användes för att beräkna k-glassoestimaten. Sedan transformerades estimatet tillbaka genom att upphöja estimatet till k. Genom 100 replikat beräknades ett medelvärde för olika utvärderingsmått. Metoden applicerades även på verklig cancerdata. Resultaten från denna studie var inte samstämmiga med originalartikelns resultat. Slutsatsen av den här studien är att metodens prestation verkar vara databeroende.

Nyckelord: *Partiell korrelation, kovariansmatris, invers, precisionsmatris, glasso, k-glasso, r-glasso, gles nätverksmodellering, SICS*

Abstract

Partial correlation between variables can be obtained implicitly through the precision matrix. To estimate the precision matrix, the sample covariance matrix can be inverted. Problems arise when the number of variables p is larger than the number of observations n, since the covariance matrix then becomes low-rank and can not be inverted. Former methods to solve this problem are uncertain and therefore a new method called *k-root-glasso* has been developed. In October 2017, the article "*Improving the Graphical Lasso Estimation for the Precision Matrix Through Roots of the Sample Covariance Matrix*" [1] was published, where the method k-glasso was presented. The claim was that the method outperforms previous methods. The aim of this study was to examine the performance of k-root-glasso on large, sparse precision matrices that are similar to networks from applications. For the simulation, two block diagonal precision matrices were generated for different values of p, where the underlying network had a *scale-free* distribution. Two variations of one of the models from the original article was also investigated. The k:th root of the empirical covariance matrix was calculated by taking the k:th root of the diagonal in its eigenvalue decomposition. The R-function *huge()* was used to calculate the estimates of k-root-glasso. Then the data was transformed back by taking the estimate to the power of k. By doing 100 replicates, the mean of the evaluation measures were computed. The method was also applied to cancer data from real observations. The results in this study were not consistent with the results from the original article. The conclusion of this study is that the performance of k-root-glasso seems to be data dependent.

Keywords: *Partial correlation, covariance matrix, precision matrix, graphical lasso, glasso, k-root glasso, r-glasso, sparse network modelling, sparse inverse covariance selection, SICS*

Innehåll

Förkortningar

λ	Straffparameter
n	Antal observationer
Ω	Precisionsmatris
$\hat{\Omega}$	Estimerad precisionsmatris
ω	Element i precisionsmatris
p	Antal parametrar
Σ	Kovariansmatris
$\hat{\Sigma}$	Estimerad kovariansmatris
S	Empirisk kovariansmatris

BIC	Bayes informationskriterium
FDR	Falsk upptäcktsfrekvens
FP	Falskt positiv
FN	Falskt negativ
MSE	Medelkvadratfel
MVN	Multivariat normalfördelning
SN	Sant negativ
SP	Sant positiv

1 Inledning

Nätverksmodeller kan användas för att beskriva komplexa system. Att estimera dessa nätverk är ett viktigt problem inom många forskningsfält såsom exempelvis medicin, biologi, och finans. Inom till exempel medicin används nätverksmodeller för att kartlägga de genetiska faktorer som ligger till grund för utveckling av olika sjukdomstyper [2]. Dessa nätverksmodeller beskriver korrelation vilket ger ett mått på hur aktiva olika gener är samtidigt. En korrelation mellan två gener kan vara missvisande om den är influerad av en tredje gen. Det är därmed ändamålsenligt att titta på korrelationen mellan två variabler, kontrollerat för alla andra variablers påverkan, så kallad partiell korrelation. Den partiella korrelationen kan implicit erhållas genom precisionsmatrisen, vilken är inversen till kovariansmatrisen. Senare i rapporten visas att maximum likelihood-estimatet för precisionsmatrisen är inversen av den empiriska kovariansmatrisen (beräknad från data).

Problem uppstår då antalet variabler, p, börjar närma sig antalet observationer, n, samt då $p > n$. Detta medför att den empiriska kovariansmatrisen får låg rang och då existerar inte dess invers. Situationen uppstår exempelvis för nätverksmodeller inom cancerforskning där det är vanligt med få observationer och många variabler. Det kan handla om ett litet antal patienter ($n \approx 200$) och ett stort antal gener ($p > 40\ 000$). För att kringgå problemet då $p > n$ tillämpas en regularisering av likelihood-funktionen som kallas 'graphical lasso' eller *glasso*. Friedman et al. [3] har utvecklat en metod som använder en blockvis optimeringsmetod för att lösa glasso-problemet. Metoden är beräkningsmässigt snabb och resulterar i ett estimat av precisionsmatrisen.

Problem som är typiska för gendata är att vissa strukturer kan bli dominanta i den empiriska kovariansmatrisen. De dominanta strukturerna i kombination med att $p > n$ resulterar i att de ledande egenvärdena blir stora och matrisrangen blir låg. För att hantera detta har en modifikation av glasso utvecklats, vid namn *k-glasso*. I oktober 2017 presenterades denna metod i artikeln "*Improving the Graphical Lasso Estimation for the Precision Matrix Through Roots of the Sample Covariance Matrix*" [1]. I k-glasso tas en rot av högre ordning, $k > 1$, ur diagonalen av den empiriska kovariansmatrisens egenvärdesdekomposition, vilket medför att de största egenvärdena dämpas. Detta föreslås leda till en mer säker estimering av precisionsmatrisen än i det otransformerade fallet.

1.1 Syfte

Syftet med detta projekt är att undersöka hur väl k-glasso presterar när det gäller estimering av stora, glesa simulerade matriser. Fokus kommer vara att jämföra glasso och k-glasso.

1.2 Problemformulering

I originalartikeln undersöks metoden k-glasso för precisionsmatriser med likformig fördelning samt med deterministisk fördelning. Hur bra fungerar k-glasso på andra typer av fördelningar som mer liknar fördelningar som påträffas inom exempelvis cancerforskning?

I originalartikelns utvärdering av metoden ligger ett stort fokus på måtten MSE och entropiförlust, som båda tar hänsyn till magnituden av länkarna i ett nätverk. I detta projekt kommer fokuset vara om k-glasso faktiskt kan estimera länkarna korrekt. Är k-glasso bättre eller sämre än glasso på att korrekt estimera länkar?

1.3 Avgränsningar

Vi kommer endast granska metoder som är baserade på maximum likelihood-estimering, och då specifikt glasso-lösaren i huge-paketet (High-Dimensional Undirected Graph Estimation) i R. Vi kommer inte fokusera på hur algoritmen glasso fungerar i detalj. Vi kommer endast att undersöka valet av optimalt λ med BIC (detta innefattar inte att fixera olika λ för jämförelse).

1.4 Etiska aspekter

I denna rapport appliceras metoden k-glasso på ett dataset bestående av observationer från hjärntumörer. Denna data är anonymiserad och utvärderingen av metoden som görs i detta projekt kommer inte att påverka enskilda individer, oavsett resultat. Utöver detta anser vi att projektet inte har några etiska aspekter att ta hänsyn till.

2 Teori

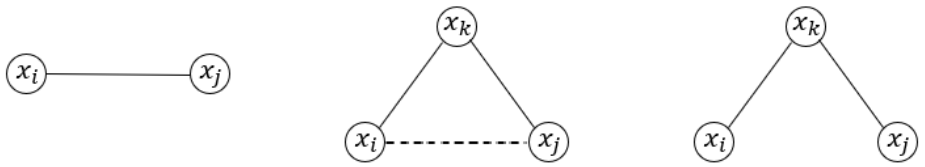
I detta avsnitt presenteras grundläggande teori om partiell korrelation, estimering av en nätverksmodell med maximum likelihood-metoden samt statistiska utvärderingsmått som behövs för att mäta ett estimats tillförlitlighet.

Grundantagandet för detta avsnitt är att det finns n observationer från multivariat normalfördelad data med p variabler, $X_{n\times p} \sim MVN(0, \Sigma = \Omega^{-1})$. Kovariansmatrisen, Σ, antas vara positivt definit och dess invers, Ω, är precisionsmatrisen. Då X är multivariat normalfördelad data är dess täthetsfunktion

$$\mathbb{P}(X = x|\Omega = \Sigma^{-1}) = \frac{1}{(2\pi)^{p/2}(\det\Omega)^{-1/2}} e^{-\frac{1}{2}x^T\Omega x}. \tag{1}$$

2.1 Partiell korrelation

Elementen i kovariansmatrisen beskriver korrelationen mellan variablerna x_i och x_j, $r_{i,j} = Cor(x_i, x_j)$, eftersom kovarians är proportionell mot korrelation. Det går dock inte att, utifrån korrelationskoefficienten, utröna om x_i och x_j faktiskt korrelerar direkt med varandra eller om korrelationen påverkas av andra variabler. Korrelationen mellan x_i och x_j skulle kunna utgöras av att en tredje variabel, x_k, är korrelerad med de båda andra variablerna, se Figur 2.



(a) (b) (c)

Figur 2: *(a) Korrelation mellan variablerna x_i och x_j. (b) Den korrelation som finns kvar mellan x_i och x_j när det kontrollerats för x_k kallas partiell korrelation (streckad linje). (c) Ingen partiell korrelation kvar givet x_k.*

Ofta är det den partiella korrelationen som är av intresse, det vill säga hur två variabler påverkar varandra direkt. Denna kan fastställas genom att beräkna korrelation mellan residualer. Residualen e_i, för en variabel x_i i linjär regression, är skillnaden mellan x_i och den delen av x_i som beskrivs av alla de variabler x_k ($k \neq i$) som regressionen utförts på. När den partiella korrelationen mellan två variabler utreds, beräknas residualerna för dessa enligt

$$e_i = x_i - \sum_{k\neq i,j} x_k\beta_k^i$$
$$e_j = x_j - \sum_{k\neq i,j} x_k\beta_k^j.$$

Om det finns kvar en korrelation mellan residualerna e_i och e_j beror det på att x_i och x_j är beroende av varandra även när all förklaringskraft som de övriga variablerna besitter har tagits bort. Det vill säga x_i och x_j är partiellt korrelerade [4]. Istället för att beräkna den partiella korrelationen direkt kan den fås genom en matrisoperation, nämligen att ta inversen av kovariansmatrisen och

därmed erhålla precisionsmatrisen. Det är bevisat att icke-diagonala element i Ω, $\omega_{i,j}$ $(i \neq j)$, är proportionella mot denna betingade korrelation givet övriga variabler [5],

$$\omega_{i,j} \propto Cor(x_i, x_j | x_k, k \neq i, j).$$

Ett förtydligande exempel är att det sägs finnas en korrelation mellan glassförsäljning och badrelaterade olyckor. Sambandet kan istället bero på en bakomliggande värmebölja som orsakar både ökad glasskonsumtion och fler badgäster. I detta exempel är alltså den partiella korrelationen mellan glassförsäljning och badrelaterade olyckor obefintlig.

2.2 Nätverk

Ett nätverk är en struktur av länkar och noder, där noder kan representera exempelvis olika gener och länkar beskriver kopplingar mellan dessa. En länk måste alltid gå mellan två noder och en nod kan ha flera länkar. Ett nätverk kan beskrivas av en så kallad *grannmatris* (eng: adjacency matrix). En grannmatris består enbart av ettor och nollor, där en etta på plats (i, j) representerar en länk mellan nod i och j, och en nolla representerar avsaknaden av en länk. Ett *glest* (eng: sparse) nätverk innebär att nätverket har få länkar.

Givet en precisionsmatris innebär $\omega_{i,j} = 0$ att variablerna i och j är betingat oberoende av varandra. En precisionsmatris kan konverteras till en grannmatris genom att sätta alla nollskilda element till 1. Därmed kan den partiella korrelationen mellan alla variabler visualiseras i ett grafiskt oriktat nätverk, där noderna är de p variablerna och avsaknaden av en länk mellan två noder innebär att de är betingat oberoende av varandra [6]. En grannmatris beskriver alltså endast förekomsten av länkar, och inte styrkan av länkarna.

I genetiska nätverk, som kommer beröras i denna rapport, motsvarar varje nod en gen och länken mellan dessa noder motsvaras av interaktionen mellan generna. Det händer att nätverken har en blockdiagonal struktur. Detta beror på att variabler som ingår i ett nätverk kan ingå i några olika separata processer. Då kommer de variabler som ingår i en process påverka varandra, men de kommer kanske inte ha någon påverkan på variabler som ingår i en annan process [7]. Dessutom är det troligt att många variabler inte påverkar varandra alls. Exempel på denna typ av nätverk kommer att presenteras i metodavsnittet.

2.3 Estimering med maximum likelihood-metoden

Precisionsmatrisen, Ω, kan estimeras från maximum likelihood-funktionen för $\Omega = \Sigma^{-1}$. Ekvation (1) medför att likelihood-funktionen för Ω är

$$L(\Omega = \Sigma^{-1}|X) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{p/2}(\det \Omega)^{-1/2}} e^{-\frac{1}{2} x_i^T \Omega x_i},$$

där x_i är en observation av p variabler. Logaritmering ger log-likelihood-funktionen

$$
\begin{aligned}
l(\Omega = \Sigma^{-1}|X) &= \sum_{i=1}^{n} \log\left(\frac{1}{(2\pi)^{p/2}(\det\Omega)^{-1/2}} e^{-\frac{1}{2}x_i^T\Omega x_i}\right) \\
&= -\sum_{i=1}^{n} \log((2\pi)^{p/2}(\det\Omega)^{-1/2}) - \frac{1}{2}\sum_{i=1}^{n} x_i^T\Omega x_i \\
&= \frac{1}{2}\sum_{i=1}^{n} \log(\det\Omega) - \frac{1}{2}\sum_{i=1}^{n} x_i^T\Omega x_i - \sum_{i=1}^{n}\frac{p}{2}\log(2\pi) \\
&= \frac{n}{2}\log(\det\Omega) - \frac{1}{2}\sum_{i=1}^{n} x_i^T\Omega x_i - C \\
&= \frac{n}{2}\log(\det\Omega) - \frac{1}{2}trace(\Omega\sum_{i=1}^{n} x_i x_i^T) - C \\
&= \frac{n}{2}\log(\det\Omega) - \frac{n}{2}trace(\Omega S) - C
\end{aligned}
$$

där C är en konstant, $S = \frac{1}{n}\sum_{i=1}^{n} x_i x_i^T$ är den empiriska kovariansmatrisen och *trace* (matrisspåret) är summan av diagonalelementen. Genom att derivera

$$\log(\det\Omega) - trace(\Omega S) \tag{2}$$

med avseende på Ω och sätta uttrycket till noll, erhålls maximum likelihood-estimatet $\hat{\Omega} = S^{-1}$.

En begränsning med detta estimat är att det får ett större systematiskt fel när förhållandet p/n närmar sig 1, samt att det inte existerar då $p > n$ eftersom S blir singulär. Vid maximering av (2) finns oändligt antal lösningar, som var för sig skulle ge en lika bra likelihood. Denna rapport kommer fokusera på en lösning från Banarjee et al. [8] som lägger till en straffparameter, λ, till (2). Parametern λ kontrollerar glesheten hos estimatet då det tvingar länkar att sättas till 0. Ett högre λ medför alltså ett glesare estimat. Detta kallas regularisering och löser problemet när $p > n$ genom att begränsa antalet lösningar av likelihooden. Den uppdaterade funktionen blir således

$$\log(\det\Omega) - trace(\Omega S) - \lambda||\Omega||_1 \tag{3}$$

där $||\Omega||_1$ är summan av absolutbeloppen av alla element i Ω. Det råder en viss förvirring över vad begreppet *graphical lasso* (hädanefter glasso) innefattar men i det här sammanhanget innebär det metoder som är fokuserade på att lösa (3). Estimatet av precisionsmatrisen är det $\hat{\Omega}$ som maximerar (3), det vill säga

$$\hat{\Omega} = \arg\max_{\Omega} \log(\det\Omega) - trace(S\Omega) - \lambda||\Omega||_1. \tag{4}$$

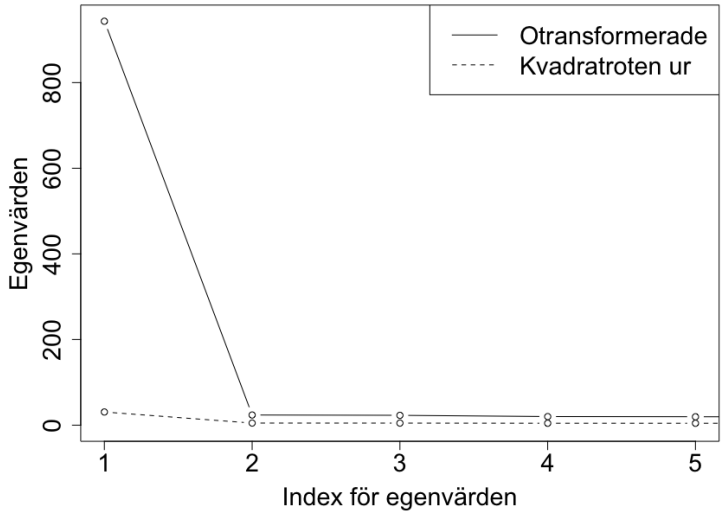
Banerjee et al. [8] visar att denna ekvation är ekvivalent med

$$
\begin{aligned}
&\hat{\Omega} = \arg\min \log(\det\Omega) \\
&\text{under villkoret} \quad ||\Omega^{-1} - S||_\infty \leq \lambda,
\end{aligned} \tag{5}
$$

där $||A||_\infty = max_{1 \leq i,j \leq p}|a_{i,j}|$. Hädanefter kommer fokus att ligga på glasso-metoden utvecklad av Friedman et al. som löser (4) genom blockvis optimering (eng: block coordinate descent) [3]. Denna är implementerad i R-paketen *glasso* och *huge*.

2.4 Metoden k-glasso

I de fall då p är stort, särskilt då $p > n$, kommer egenvärdena till S bli mer spridda [9] och de minsta egenvärdena underskattas medan de största överskattas [1]. När kvoten mellan p och n ökar, så ökar även konditionstalet [10]. Det stora konditionstalet innebär att alla beräkningar som involverar S blir numeriskt instabila och får systematiska fel. Alltså leder $p > n$ till att glasso-estimatet blir instabilt. De största egenvärdena beskriver den största variansen hos en matris. Eftersom egenvärdena för ett block också är egenvärden för hela matrisen [11], kommer de största av dessa för blockdiagonala matriser att till stor del beskriva blockstrukturen. Därför kan glasso missa vad som händer inom och utanför block där variansen är liten i förhållande till blocken. För att få ett mer exakt estimat föreslår Avagyan et al. [1] i en artikel från 2017 en metod som benämns *k-root-glasso*, hädanefter k-glasso.



Figur 3: *De fem största egenvärdena för S då $p = 200$ och $n = 80$. Roten $k = 2$ dämpar de största egenvärdena markant.*

Antag att egenvärdesdekompositionen av $S = PDP'$. Den k:te roten av S definieras som $S^{1/k} = PD^{1/k}P'$, där $k > 1$. Detta minskar spridningen av S egenvärden eftersom stora egenvärden blir mindre och egenvärden mindre än 1 blir större. Se Figur 3 för en jämförelse av egenvärdesspridningen. Metoden k-glasso innebär att lösa glasso-problemet (5) med $S^{1/k}$ för att estimera $\hat{\Omega}_{k-glasso}$. Därefter transformeras estimatet tillbaka genom att upphöjas med k, det vill säga $\hat{\Omega} = \hat{\Omega}^k_{k-glasso}$. Notera att $k = 1$ är ekvivalent med glasso. Denna transformation av S leder enligt Avagyan et al. [1] till en mer tillförlitlig skattning av Ω än vad estimering utan transformation skulle göra. De föreslår därmed det från (5) uppdaterade problemet

$$\hat{\Omega}_{k-glasso} = \arg\min \log(\det \Omega)$$
$$\text{under villkor} \quad ||\Omega^{-1/k} - S^{1/k}||_\infty \leq \lambda,$$

där λ är straffparametern. Vid högdimensionell data innehåller $\hat{\Omega}_{k-glasso}$ inga nollelement utan endast element som är nära noll. Avagyan et al. menar att estimatet är approximativt glest. För att handskas med detta problem används en tröskelregel, vilket innebär att alla element vars absolutbelopp är mindre än ett tröskelvärde (förslagsvis 10^{-10} enligt Avagyan et al.) sätts till noll [1]. Det är viktigt att tillämpa tröskelregeln efter att estimatet transformerats tillbaka till originalskalan.

Avagyan et al. [1] kommer fram till att k-glasso presterar bättre än glasso när det gäller statistiska förluster, specificitet och sensitivitet, i synnerhet för nätverksmodeller med deterministisk fördelning. Deras studie visar att roten $k = 2$ i genomsnitt presterar bäst.

2.4.1 Val av λ

Valet av λ är viktigt eftersom denna kontrollerar skattningens gleshet. Avagyan et al. [1] föreslår användandet av en variant av *Bayes informationskriterium* (eng: Bayesian Information Criterion, hädanefter BIC) definierad av Yuan et al. [12] för att välja λ,

$$BIC(\lambda) = n(-\log\det(\hat{\Omega}(\lambda)) + trace(S\hat{\Omega}(\lambda))) + \log(n) * NZ, \tag{6}$$

där NZ är antalet nollskilda element i $\hat{\Omega}$. Givet en sekvens av λ-värden väljs det λ med tillhörande estimat, $\hat{\Omega}(\lambda)$, som ger det lägsta värdet på (6). Det ska dock noteras att BIC är baserad på likelihood och väljer därför modell utifrån anpassningsgrad (eng: goodness of fit) och inte utifrån exempelvis stabilitet av modellselektion. Andra sätt att välja λ för att erhålla ett optimalt estimat är till exempel genom korsvalidering.

2.5 Utvärdering av estimat

Utvärdering av simulerad data innebär att jämföra det estimerade nätverket med det sanna nätverket. Jämförelsen kan gå till på många olika sätt beroende på vad som anses vara viktigt. Tre typer av mått kommer att användas i denna rapport: *mått baserade på klassificering*, *likelihood-baserade mått* samt *topologiska mått*.

2.5.1 Mått baserade på klassificeringstabellen

I klassificeringstabellen, Tabell 1, visas andelen *falskt positiva* (FP), *sant positiva* (SP), *falskt negativa* (FN) och *sant negativa* (SN) länkar i estimatet. En sant respektive falskt positiv länk är en estimerad länk som finns respektive inte finns i det sanna nätverket. En sant respektive falskt negativ länk är en estimerad icke-länk som inte finns respektive finns i det sanna nätverket. Det är förstås fördelaktigt om diagonalerna i Tabell 1 (antalet SN respektive SP) dominerar storleksmässigt.

Tabell 1: *Klassificeringstabell - fördelning av antal sant negativa (SN), falskt negativa (FN), falskt positiva (FP) och sant positiva (SP) länkar i det estimerade nätverket.*

		Sant nätverk	
		0	1
Estimerat nätverk	0	# SN	# FN
	1	# FP	# SP

Specificitet är andelen korrekt estimerade icke-länkar av alla icke-länkar i det sanna nätverket: $\frac{SN}{SN+FP}$. Det är fördelaktigt att denna kvot går mot 1 eftersom det innebär att metoden har estimerat ett lågt antal falska länkar. *Sensitivitet* mäter andelen korrekt estimerade länkar i förhållande till totala antalet länkar i det sanna nätverket: $\frac{SP}{SP+FN}$. På samma sätt som för specificitet är det även här fördelaktigt om kvoten går mot 1 [13]. Dessa mått tenderar dock att vara intetsägande när det finns obalanserade grupper i datan [14].

Ett mått som beskriver mängden av alla felaktigt estimerade länkar av alla länkar i estimatet är *falsk upptäcktsfrekvens* (eng: false discovery rate, hädanefter FDR). Denna beräknas som $\frac{FP}{FP+SP}$. Det är därmed eftersträvansvärt att FDR blir så låg som möjligt eftersom estimatet i sådant fall estimerar ett lågt antal falska länkar [15]. Ibland kan det vara enklare att istället använda måttet *positivt prediktionsvärde* (eng: positive predictive value, hädanefter PPV). Detta är komplementet till FDR och kan således beräknas som $1-\text{FDR}$ [13].

2.5.2 Likelihood-baserade mått

Medelkvadratfelet (eng: mean squared error, hädanefter MSE) för ett estimat av Ω är $E[(\Omega - \hat{\Omega})^2]$. Måttet är ekvivalent med likelihood för normalfördelad data, med en skalfaktor och förskjutning [16]. Detta betyder att MSE är ett mått på hur väl modellen stämmer överens med observationerna. Om intresset i skattningen av ett nätverk är att ta reda på om variabler påverkar varandra eller inte, är anpassningsgraden av mindre betydelse. MSE tar endast hänsyn till magnituden av skillnaden mellan element i estimat och sant nätverk och inte huruvida en länk finns eller inte finns.

2.5.3 Topologiska mått

Till skillnad från ovan presenterade mått lägger topologiska mått större vikt på samstämmigheten i vilka noder som på olika sätt är viktiga för nätverket.

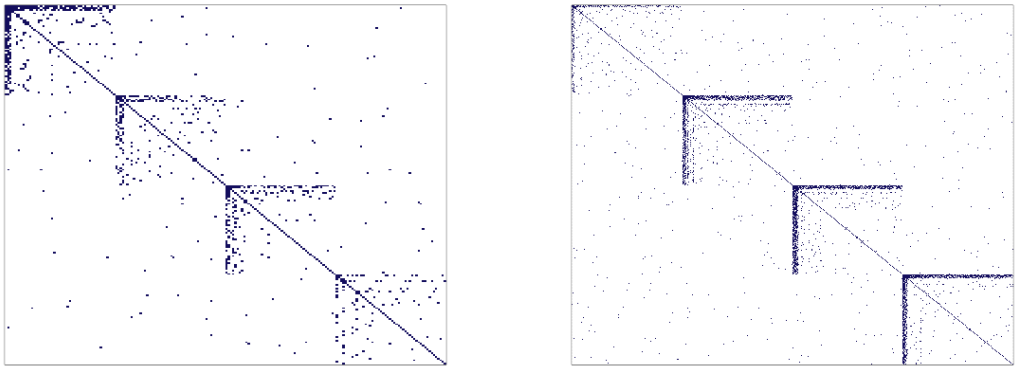
Centralitet (eng: centrality) är ett mått som mäter hur central eller inflytelserik en nod är i ett nätverk. För en nod i, definieras dess *gradcentralitet* (eng: degree centrality) som antalet intilliggande noder som är direkt länkade till i. Det innebär alltså att gradcentraliteten anger hur viktig en nod är i ett nätverk. *Sammankopplingscentralitet* (eng: betweeness centrality) är ett mått som definieras av hur många gånger en nod passeras i de kortaste stiglängderna (eng: pathlength) mellan två andra noder i nätverket [17].

3 Metod

För att utvärdera k-glasso behövs data där det finns ett sant underliggande nätverk. I denna studie användes data simulerad från fyra underliggande sanna precisionsmatriser Ω. Två av precisionsmatriserna genererades särskilt för denna studie. De övriga var en precisionsmatris från artikeln av Avagyan et al. [1] samt en modifierad version av densamma. Vidare i detta avsnitt beskrivs hur k-glasso tillämpats på den genererade datan för att få ett estimat $\hat{\Omega}$ och hur valet av k och λ gick till. Utöver detta applicerades även metoden på data för hjärntumörer. All simulering genomfördes i *R* och koden återfinns i Appendix Bilaga E.

3.1 Generering av sanna precisionsmatriser Ω

Med hjälp av R-funktionen *barabasi.game()* från paketet *igraph* genererades fyra glesa grannmatriser med *scale-free* fördelning. Denna fördelning valdes då den påstås efterlikna genetiska nätverk [18]. En blockdiagonal matris skapades, där de fyra genererade matriserna placerades på diagonalen. För att översätta grannmatriserna till precisionsmatriser ersattes ettorna i grannmatriserna med värden från en normalfördelning. Därefter slumpades $0.0025p^2$ normalfördelade länkar fram utanför blocken. Detta gjordes för $p = 200$ och $p = 500$ och resulterade i de två sanna precisionsmatriserna *Prec SF200* respektive *Prec SF500*. De två tillhörande grannmatriserna går att se i Figur 4. Valet av p styrdes av intresset att få ett enklare, lättöverskådligt fall, samt ett fall med så många variabler som möjligt utan att det blev för beräkningsintensivt.



(a) *Prec SF200 med gleshet = 0.973*

(b) *Prec SF500 med gleshet = 0.980*

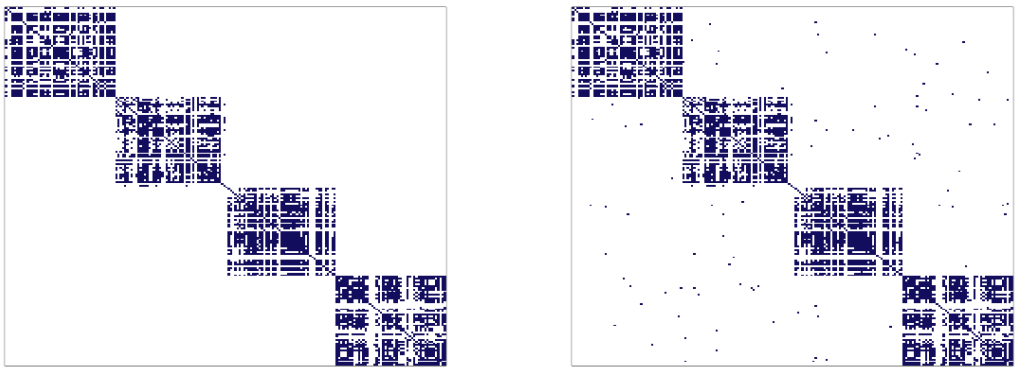
Figur 4: *De simulerade grannmatriserna.*

Precisionsmatrisen som användes från studien av Avagyan et al. [1] var den som benämns 'Model 4' med $p = 200$ (hädanefter *Prec U200*). Även den hade fyra block längs diagonalen. Till skillnad från de ovan nämnda precisionsmatriserna var blocken i Prec U200 betydligt tätare och det fanns inga element utanför blocken. Länkarna var likformigt fördelade inom blocken och länkarnas värden var normalfördelade. För den modifierade versionen slumpades $0.0025p^2$ normalfördelade länkar fram utanför blocken (hädanefter *Prec U200+*). De två tillhörande grannmatriserna går att se i Figur 5.

När länkar lades till utanför blocken i precisionsmatriserna blev matriserna inte längre inverterbara, vilket de behöver vara för att kunna generera data. För att reparera detta ökades magnituden på länkarna i diagonalen med absolutbeloppet av det minsta egenvärdet plus en konstant, för att

egenvärdena skulle bli nollskilda.

Elementen i precisionsmatriserna som var mindre än tröskelnivån 10^{-10} sattes till noll.



(a) *Prec U200, gleshet = 0.879*

(b) *Prec U200+, gleshet = 0.876*

Figur 5: *Grannmatrisen från artikeln av Avagyan et al. samt dess modifierade version.*

3.2 Simulering via k-glasso

För var och en av de fyra precisionsmatriserna Ω genererades data $X_{n\times p} \sim MVN(0, \Sigma = \Omega^{-1})$ med $n = 0.4p$ respektive $n = 5p$ observationer, där datan hade samma korrelationsstruktur som Ω. Från den genererade datan beräknades den empiriska kovariansmatrisen S. För ett givet k beräknades $S^{1/k}$ enligt definitionen i Avsnitt 2.4. Konditionstalet för $S^{1/k}$ beräknades som kvoten mellan det högsta och det minsta nollskilda egenvärdet, där toleransen sattes till 10^{-10}. R-funktionen *huge()* användes för att beräkna k-glassoestimatet $\hat{\Omega}$, för en sekvens av λ-värden. Från denna sekvens valdes ett λ ut på tre olika sätt som beskrivs nedan. Detta upprepades 100 gånger för att sedan kunna beräkna medelvärden för utvärderingsmåtten.

3.2.1 Undersökta värden på k

I artikeln från 2017 föreslog Avagyan et al. att k skulle väljas med hjälp av BIC. Dock observerade de att en förvald rot $k = 2$, skulle vara ett bra val på k. I deras exempelsimulering går det att observera att BIC nästan alltid minimeras av ett k i närheten av, och oftast under, 2 [1]. Därför jämfördes metoderna i denna studie med tre inställningar av k; $k = 1$ (original glasso), $k = 1.5$, samt $k = 2$.

3.2.2 Val av λ-sekvens

Sekvensen av λ valdes så att tillhörande estimat skulle täcka ett intervall av gleshet omkring det sanna nätverkets gleshet. Eftersom k-glasso innebär en transformation av S är betydelsen av λ förändrad i glasso-lösaren. Därmed behöver även λ transformeras så att denna betyder samma sak även när värdena på k varierar. En bra transformation av λ skulle alltså innebära att λ_i för exempelvis $k = 1$, genererade ett nätverk av samma gleshet som λ_i' för $k = 2$. Sekvenserna valdes för varje kombination av n, k och nätverk, genom att iterativt undersöka olika transformationer för att uppnå denna matchning.

3.2.3 Val av λ

För varje k valdes ett λ ur respektive λ-sekvens och tillhörande $\hat{\Omega}$ på tre olika sätt. Dels valdes λ ut via BIC, se Ekvation (6). Dels valdes det λ ut, vars estimat av Ω hade en gleshet som var närmast det sanna nätverkets gleshet. Slutligen valdes det λ ut, vars estimat av Ω hade en FDR som var närmast en fördefinierad önskad FDR på 0.2. För både matchningen av FDR och gleshet, betraktades det utvalda estimatet som giltigt om dess FDR eller gleshet var inom ± 0.02 från det sökta värdet. Om mer än 80% av alla replikat hade ogiltiga estimat vid matchning mot FDR respektive gleshet ansågs den matchningen vara ogenomförbar. När BIC ska beräknas numeriskt kan det hända att även om log-determinanten av $\hat{\Omega}$ existerar, blir determinanten av $\hat{\Omega}$ oändligt stor. Detta löses genom att istället beräkna summan av logaritmen av alla egenvärden, eftersom determinanten är ekvivalent med produkten av alla egenvärden.

3.3 Utvärdering

Estimaten utvärderades utifrån måtten specificitet, senitivitet, FDR, MSE samt andel noder med högst centralitetsmått (grad- respektive sammankopplingscentralitet). Samstämmigheten i centralitet jämfördes genom att se hur stor andel av de viktigaste noderna i ett estimat, som överensstämmer med de viktigaste noderna i det sanna nätverket. Till exempel, om alla noder har unika centralitetsmått, både i estimatet och i det sanna nätverket, är topp 10% av 200 variabler ett antal om 20 noder. Om 5 av dessa matchar topp 10% av de största noderna i det sanna nätverket, innebär det en matchning på $5/20 = 0.25$. Även ett mått på en proportion på var metoderna lägger de falskt positiva (FP) länkarna, innanför eller utanför blocken, estimerades.

Måtten FDR och sensitivitet presenteras tillsammans för att se förhållandet mellan andelen felaktigt estimerade länkar och andelen korrekt estimerade länkar. Detta gjordes genom att plotta FDR mot sensitivitet för varje λ i en λ-sekvens.

3.4 Observerad data från hjärntumörer

Slutligen testades k-glasso på ett dataset innehållandes 508 observationer av 2371 transkriptionsfaktorer [1] från hjärntumörer. Eftersom detta är observerad data fanns inte någon sann precisionsmatris att utvärdera estimatet mot. Däremot finns listor över länkar, baserat på vad som är bevisat på experimentell nivå, över vilka gener som observerats interagera med varandra. Interaktionen tros inte nödvändigtvis vara statisk, utan den ska ses som approximativ och alltså ge en indikation på vilka länkar som borde finnas i nätverket. Utvärderingen bestod således enbart i att jämföra hur många av dessa länkar som estimaten innehöll.

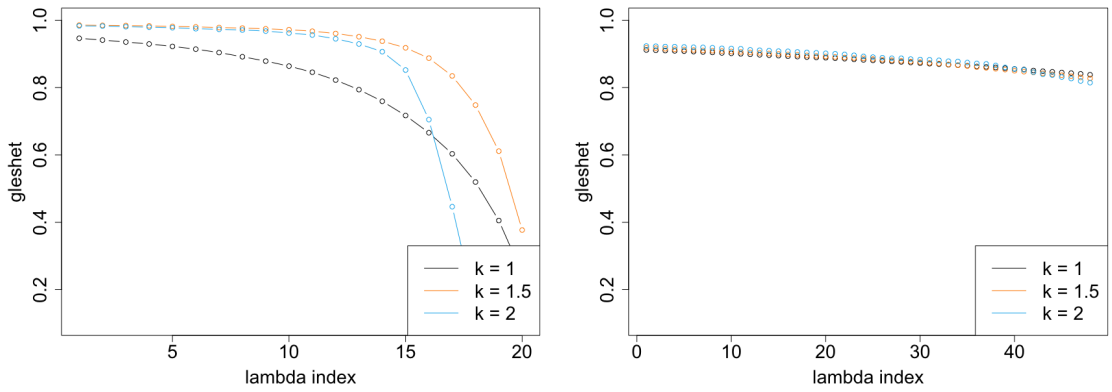
[1] Transkriptionsfaktorer är en typ av protein som reglerar vilka gener som är aktiva i en viss cell. Denna reglering påverkar celltillväxten och kan orsaka mutationer om regleringsförmågan inte fungerar som den ska. Processen har betydelse för celltransformation och vid vissa sjukdomstyper har ovanligt låga eller höga värden på vissa transkriptionsfaktorer observerats. Att identifiera dessa transkriptionsfaktorer kan vara en viktig del inom till exempel cancerforskning [19].

4 Resultat

I detta avsnitt presenteras resultaten från simuleringarna. Inledningsvis introduceras en jämförelse mellan denna studies val av λ-sekvens och den λ-sekvens som återfinns i Avagyan et al. [1] Vidare presenteras utvärderingsmåtten för den simulerade datan samt datan från hjärntumörer, först då λ valdes med BIC och sedan då valet av λ gjordes för att matcha det ursprungliga nätverkets gleshet. Eftersom få estimat kunde uppfylla den valda nivån av FDR redovisas inte dessa resultat. I de tabeller som finns återges tal med tre signifikanta siffror. I avsnittet presenteras figurer över utvärderingsmått, men detaljerade tabeller återfinns i Appendix. Resultatet av MSE presenteras inte i figurerna utan återfinns enbart i Appendix.

4.1 Val av λ-sekvens

Glesheten hos Prec SF200 och Prec SF500 var cirka 0.98 och då konstruerades en sekvens av λ som täckte en gleshet från cirka 0.9 till 1. Hos Prec U200 och Prec U200+ låg glesheten på cirka 0.87 och där konstruerades en sekvens av λ som täckte en gleshet från cirka 0.8 till 0.95. I artikeln av Avagyan et al. [1] föreslogs en sekvens av λ, som för deras nätverksmodeller täckte en estimering från ett nästintill tomt till ett nästan fullt nätverk. En jämförelse mellan artikelns λ-sekvens och sekvensen som valdes i detta projekt illustreras i Figur 6. Vi kan se att även om k-glasso med $k = 1$ använder det högsta möjliga λ kan det inte estimera ett lika glest nätverk som k-glasso med övriga rötter.



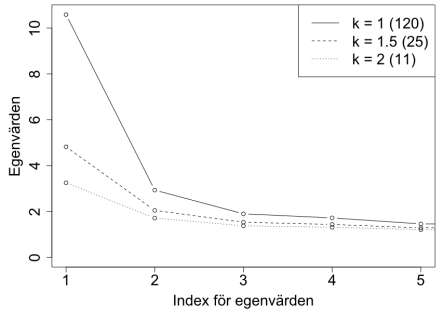
(a) *Sekvenser från Avagyan et al.*

(b) *Sekvenser konstruerade för denna rapport*

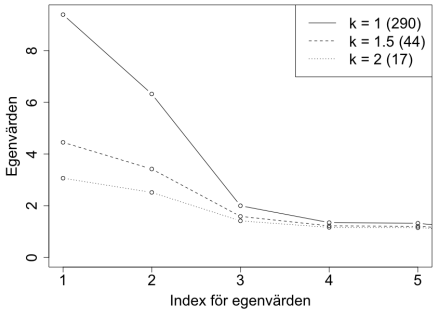
Figur 6: *Gleshet för estimat av nätverksmodellen Prec U200 över en avtagande sekvens av λ. Det sanna nätverket hade en gleshet på 0.87. (a) Sekvenser av λ som Avagyan et al. använde i sin simulering. Sekvensen täckte estimering från ett nästan tomt till ett nästan fullt nätverk. Sekvenserna för glasso med $k > 1$ estimerar genomgående glesare nätverk. (b) Sekvenser av λ för samma nätverksmodell som användes i denna simulering. Sekvenserna täcker gleshet av estimat för ett intervall kring det sanna nätverkets gleshet. Ett λ_i för ett visst k motsvarar ett estimat av ungefär samma gleshet som ett λ_i för de övriga värdena på k.*

4.2 Egenvärdesspridning av den genererade datan

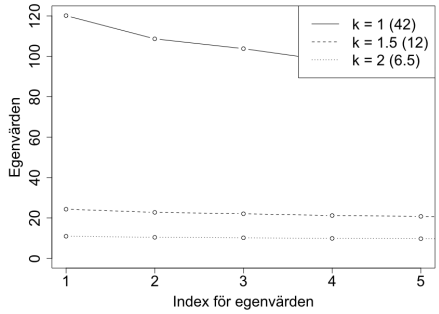
I Figur 7 ses egenvärdesspridningen för de fem största egenvärdena samt konditionstal för empiriska kovariansmatrisen $S^{1/k}$ då $p > n$. Det största egenvärdet är betydligt högre för Prec U200 och Prec U200+ än de övriga två.



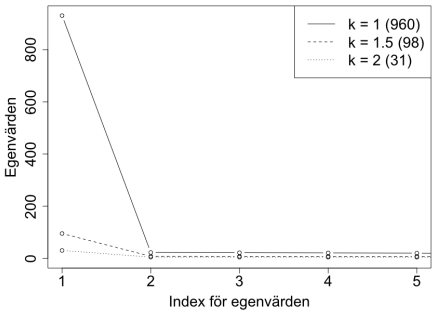
(a) Prec SF200



(b) Prec SF500



(c) Prec U200



(d) Prec U200+

Figur 7: *Egenvärdesspridning för de fem största egenvärdena hos S efter generering av data då $p > n$ från de olika nätverken. Transformation med $k = 2$ reducerar de högsta egenvärdena avsevärt. Notera att dimensionen för y-axeln skiljer sig åt mellan nätverksmodellerna. I figurerna syns även konditionstalen för $S^{1/k}$ inom parentes.*

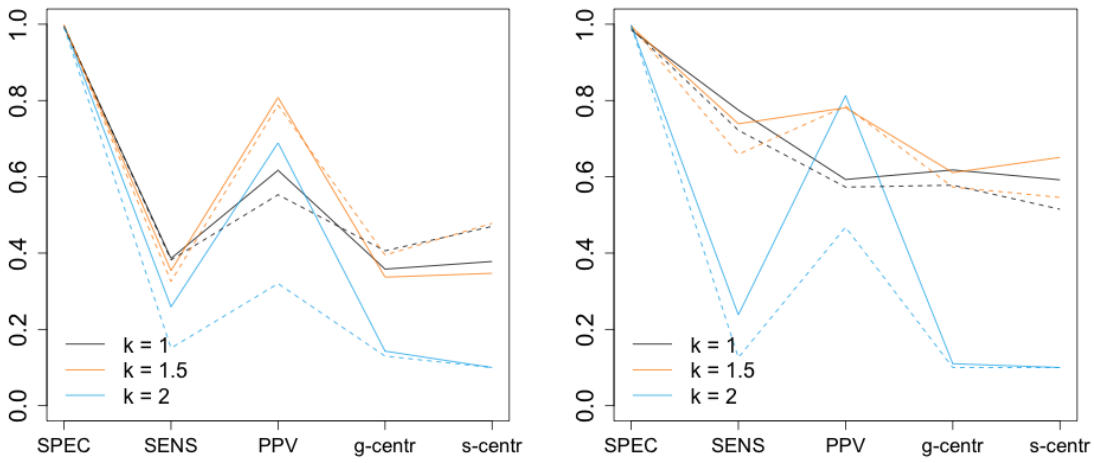
4.3 Utvärdering av simulerad data

Nedan redovisas utvärdering av de 100 replikaten. Hädanefter innebär benämning av centralitetsmått den matchning av topp 10 % noder som nämns i Avsnitt 3.3. I de figurer som visar en sammanställning av de olika utvärderingsmåtten innebär höga värden önskvärda resultat.

4.3.1 Val av λ baserat på BIC: Prec SF200 och Prec SF500

I Figur 8 presenteras resultaten från de olika utvärderingsmåtten. Vi kan se att det inte finns några nämnvärda skillnader i specificitet för de olika k-rötterna. Roten $k = 2$ ger generellt lägre värden för de övriga måtten. Undantaget är värdet på PPV då Prec SF200 med $k = 2$ resulterar i högst värde när $p < n$. Notera att estimatet för $k = 2$ är glesare än estimaten för övriga k, se Tabell 2.

Fördelningen av falskt positiva (FP) länkar visar att $k = 2$ lägger en övervägande majoritet av alla FP länkar inom blocken medan de övriga rötterna lägger sina FP länkar utanför, se Tabell C1 i Bilaga C. Proportionen av FP länkar som placeras inom eller utanför block kan vara missvisande då den inte tar hänsyn till att det totala antalet FP länkar kan vara av helt olika storleksordningar för olika k. Vi kan dock se att $k = 2$ placerar sina FP länkar inom block både när det har sämre och bättre FDR än de övriga.



(a) $p > n$ (b) $p < n$

Figur 8: *De heldragna respektive streckade linjerna visar utvärderingsmått för Prec SF200 respektive Prec SF500 då λ valdes med BIC.*

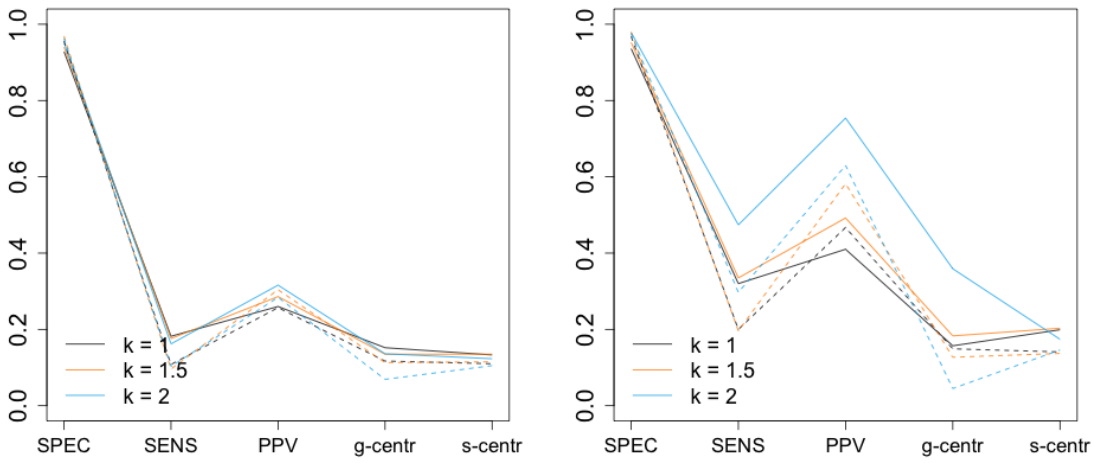
Resultaten för $k = 1$ och $k = 1.5$ följs åt för de flesta utvärderingsmåtten. Roten $k = 1$ estimerar genomgående tätare nätverk än övriga k och i fallet $p < n$ estimerar $k = 2$ glesare nätverk än de övriga, se Tabell 2. Vad gäller MSE följer detta mått inte trenden då $k = 2$ får lägst MSE då $p > n$.

Tabell 2: *Gleshet för estimaten av Prec SF200 respektive Prec SF500.*

	p > n		p < n	
	Prec SF200	Prec SF500	Prec SF200	Prec SF500
k = 1	0.983	0.986	0.965	0.975
k = 1.5	0.988	0.992	0.975	0.983
k = 2	0.990	0.991	0.992	0.995

4.3.2 Val av λ baserat på BIC: Prec U200 och Prec U200+

I Figur 9 visualiseras resultaten för Prec U200 och Prec U200+. För $p > n$ kan vi inte se några stora skillnader varken för de olika värdena på k eller de två nätverken. När $p < n$ ser vi att $k = 2$ resulterar i något högre värden, i synnerhet för PPV.



(a) $p > n$ (b) $p < n$

Figur 9: *De heldragna respektive de streckade linjerna visar utvärderingsmått för Prec U200 respektive Prec U200+ då λ valdes med BIC.*

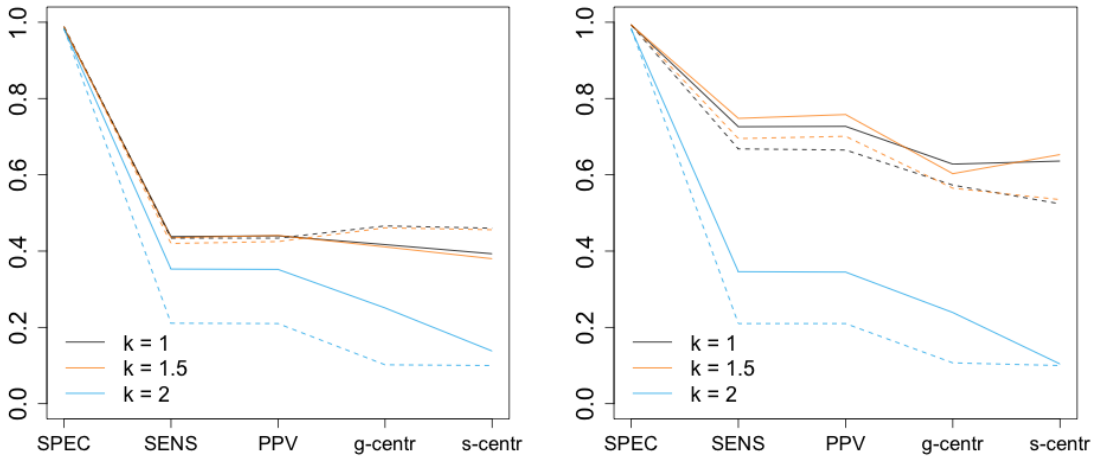
Återigen noterar vi att val med BIC för $k = 2$ i regel har estimerat ett glesare nätverk, vilket kan påverka måtten, se Tabell 3. Andelen FP inom block är högst för $k = 2$ som lägger majoriteten av sina FP länkar inom block, utom för Prec U200 då $p > n$, se Tabell C2 i Bilaga C. Det går inte att utvärdera MSE då den är så pass liten att den avrundas till noll med tre decimalers noggrannhet.

Tabell 3: *Gleshet för estimaten av Prec U200 respektive Prec U200+.*

	p > n		p < n	
	Prec SF200	Prec SF500	Prec SF200	Prec SF500
k = 1	0.915	0.948	0.905	0.947
k = 1.5	0.926	0.961	0.917	0.958
k = 2	0.937	0.955	0.924	0.942

4.3.3 Val av λ baserat på gleshet: Prec SF200 och Prec SF500

I Figur 10 kan vi se samma tendenser som med val genom BIC, det vill säga att $k = 1$ och $k = 1.5$ tenderar att följa samma mönster medan $k = 2$ ger markant lägre värden för alla mått utom specificitet. Vi ser här att när nätverken tvingas vara av samma gleshet får $k = 2$ högre PPV. När det gäller andelen FP länkar inom och utanför block lägger $k = 2$ även här störst andel FP länkar inom blocken, medan det motsatta gäller för $k = 1$ och $k = 1.5$ (se Tabell C3 i Bilaga C). Det finns ingen genomgående trend i vilket k som får lägst MSE.

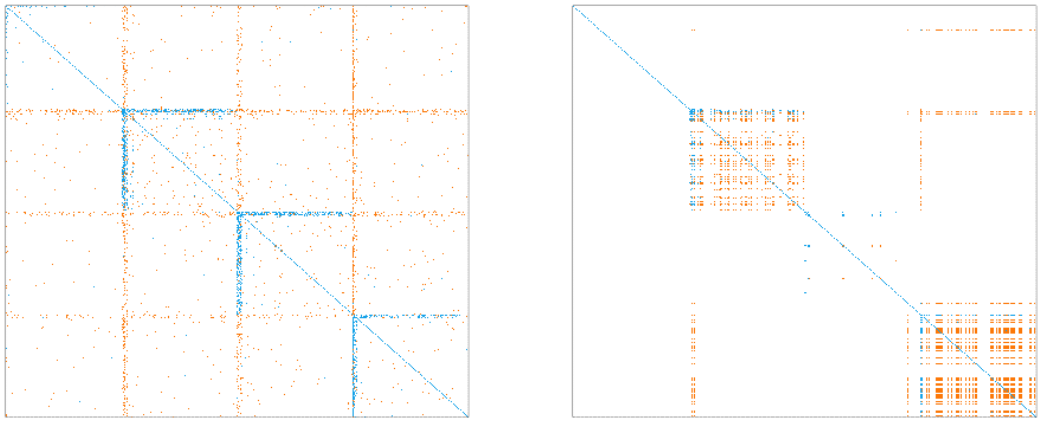


(a) $p > n$ (b) $p < n$

Figur 10: *De heldragna respektive streckade linjerna visar utvärderingsmått för Prec SF200 respektive Prec SF500 då λ valdes genom en matchning av gleshet.*

Vid jämförelse av FDR och sensitivitet bekräftas resultatet att $k = 2$ är ett, i vissa fall avsevärt, sämre val för den simulerade datan, samt att $k = 1$ och $k = 1.5$ är ungefär likvärdiga, se Bilaga D i Appendix.

I Figur 11 visualiseras hur länkarna i estimatet av Prec SF500 är fördelade för två olika k, jämfört med det sanna nätverket (se Figur 4b). För $k = 1$ är de korrekt estimerade länkarna (blå färg) placerade i en struktur som tydligt påminner om den sanna strukturen. För $k = 2$ försvinner denna struktur eftersom många falska länkar (orange färg) estimeras och bildar två täta block.



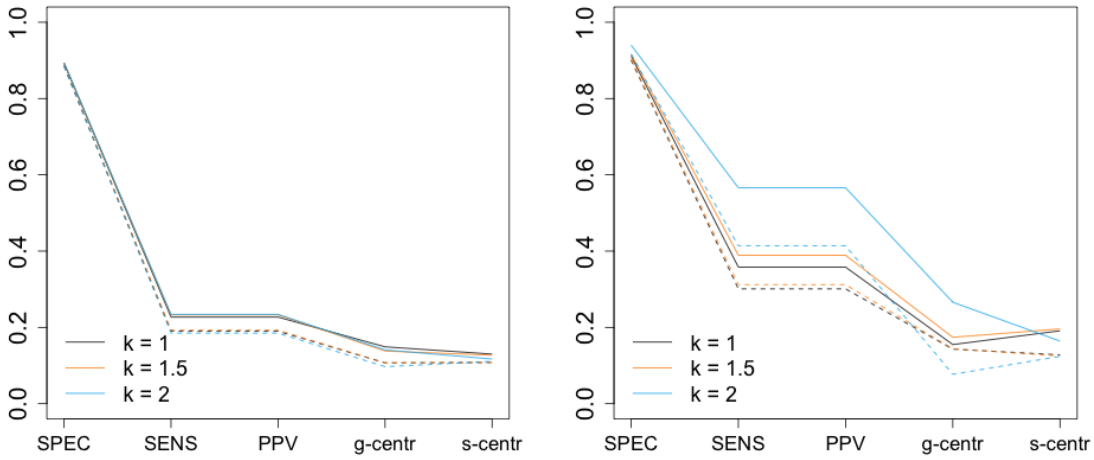
(a) $k = 1$ (b) $k = 2$

Figur 11: *Grannmatriser från estimat av Prec SF500, valda med gleshet motsvarande den sanna glesheten. En blå punkt är en sant positiv länk och en orange punkt är en falskt positiv länk.*

4.3.4 Val av λ baserat på gleshet: Prec U200 och Prec U200+

Figur 12 visar resultaten för Prec U200 och Prec U200+. I fallet $p > n$ ser vi att det inte finns några större skillnader mellan olika k, men $k = 2$ presterar bättre än de övriga då $p < n$. Vid $p > n$ finns det inga större skillnader mellan olika k gällande var de flesta FP länkarna placeras, de flesta läggs utanför blocken (se Tabell C4 i Bilaga C). I fallet $p < n$ lägger $k = 2$ något fler FP länkar inom block, men vi noterar att denna rot samtidigt får högst PPV. Även här är MSE så pass litet att det inte går att jämföra.

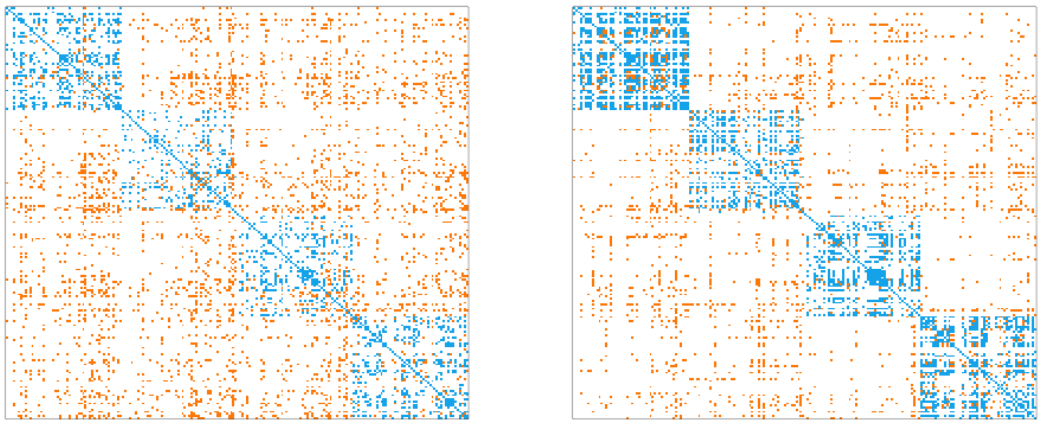
När vi jämför FDR och sensitivitet förstärks bilden av att $k = 2$ är bättre än övriga k då $p < n$ och att skillnaden minskar då $p > n$, se Bilaga D.



(a) $p > n$ (b) $p < n$

Figur 12: *De heldragna respektive streckade linjerna visar utvärderingsmått för Prec U200 respektive Prec U200+ då λ valdes genom en matchning av gleshet.*

I Figur 13 visualiseras hur länkarna i estimatet av Prec U200+ är fördelade för två olika k, jämfört med det sanna nätverket (se Figur 5b). I fallet då $k = 1$ är det svårt att utröna någon tydlig struktur av estimerade länkar. Inom blocken är de flesta länkar SP men vi noterar att det sanna nätverket också har en tät fördelning av länkar inom blocken. För $k = 2$ blir blockstrukturen tydligare då färre länkar utanför blocken estimerats, medan fler länkar inom block är korrekt estimerade som sanna länkar.

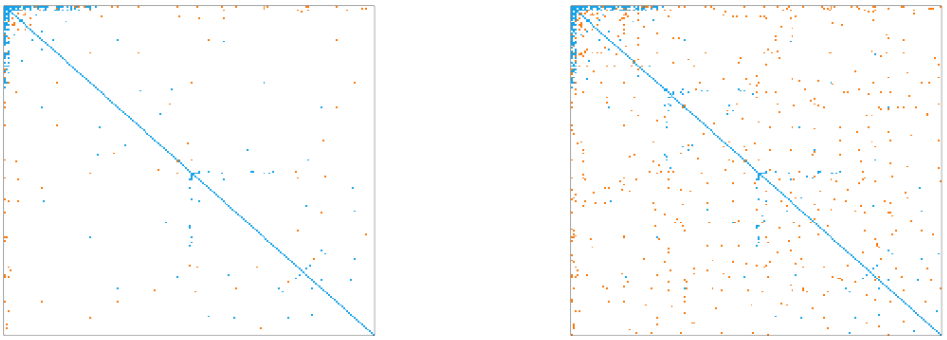


(a) $k = 1$ (b) $k = 2$

Figur 13: *Grannmatriser för estimat av Prec U200+, valda med gleshet motsvarande den sanna glesheten. En blå punkt är en sant positiv länk och en orange punkt är en falskt positiv länk.*

4.3.5 En jämförelse mellan estimat av olika gleshet

I Figur 14 illustreras en jämförelse mellan estimat av samma nätverk, men med olika gleshet. Den första grannmatrisen är ett estimat med *gleshet* $= 0.988$ där λ valdes via BIC, medan den andra grannmatrisen är ett estimat med *gleshet* $= 0.974$ och där λ valdes genom att fixera glesheten. Den sistnämnda är alltså mindre gles och måste estimera fler länkar, varav de flesta länkar blir falska (orange färg).

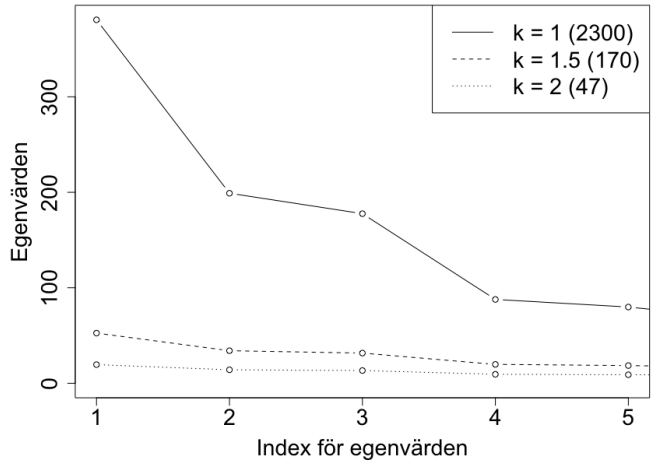


(a) *Estimat med gleshet* $= 0.988$ (b) *Estimat med gleshet* $= 0.974$

Figur 14: *Grannmatriser av estimat för Prec SF200,* $n = 80$*,* $k = 1.5$*. I (a) valdes* λ *via BIC och i (b) valdes* λ *genom att fixera glesheten till vald nivå. En blå punkt är en sant positiv länk och en orange punkt är en falskt positiv länk.*

4.4 Utvärdering av k-glasso på data från hjärntumörer

Egenvärdesspridning och konditionstal för $S^{1/k}$ från observationerna av transkriptionsfaktorer illustreras i Figur 15. Konditionstalen är högre än för den simulerade datan och transformationerna dämpar de största egenvärdena.



Figur 15: *De fem största egenvärdena från data av hjärntumörer. Transformationerna* $k = 1.5$ *och* $k = 2$ *dämpar de största egenvärdena. Konditionstalen för* $S^{1/k}$ *redovisas inom parentes.*

Estimaten utvärderades utifrån listan av kända länkar för nätverket. Hur stor andel av de kända länkarna som återfinns hos estimaten, redovisas som SENS*. Måttet beräknas på samma sätt som sensitivitet men på grund av avsaknaden av ett sant nätverk poängteras skillnaden genom denna benämning. Andelen korrekt estimerade kända länkar av totala antalet estimerade länkar anges av samma anledning som PPV*.

Tabell 4 innehåller resultat då λ valts med BIC. Vi noterar att estimaten är de samma som i Tabell 5 förutom estimatet av k-glasso med $k = 2$ som är något glesare. Detta estimat får avsevärt högre PPV*, vilket innebär att av de länkar som är estimerade är majoriteten sanna länkar.

I Tabell 5 redovisas resultat från estimering av transkriptionsfaktorerna då λ valts utifrån en gleshetsnivå inom intervallet $[0.992, 0.999]$. Vi ser att estimaten har ungefär samma SENS* men att estimatet för k-glasso med $k = 1$ har ett lägre PPV* än de övriga.

Tabell 4: *Estimat då* λ *valts med BIC.*

	SENS*	PPV*	Gleshet
k = 1	0.148	0.309	0.999
k = 1.5	0.148	0.514	0.999
k = 2	0.146	0.855	1.000

Tabell 5: *Estimat då* λ *valts utifrån en gleshet inom* $[0.992, 0.999]$.

	SENS*	PPV*	Gleshet
k = 1	0.148	0.309	0.999
k = 1.5	0.148	0.514	0.999
k = 2	0.149	0.541	0.999

5 Diskussion

Inledningsvis diskuteras upplägget av simuleringsstudien för att sedan gå över till att diskutera resultatens betydelse.

5.1 Betydelsen av λ-sekvens

Att hitta rätt λ-sekvens är ett dataspecifikt problem. Det fanns vissa mönster i vilka transformationer som behövde göras för att sekvenserna för olika värden på k skulle generera estimat av samma gleshet. Dessa mönster var dock inte konsekventa nog för att samma regel skulle kunna överföras till vare sig olika nätverksmodeller eller förhållandet mellan antal variabler och antal observationer. Detta var en tidskrävande del av simuleringen. Eftersom det i framtida simuleringar skulle vara intressant att iterera över fler värden på k är detta ett problem att ta ställning till. Därför skulle det vara givande att matematiskt kunna härleda ett samband mellan valet av λ för olika k, om ett sådant existerar. Enligt Figur 6a använde Avagyan et al. [1] λ-sekvenser som hade olika täckning av gleshet för olika k. Därmed fanns en möjlighet att spridningen av gleshet mellan de olika estimaten blev stor, i synnerhet eftersom valet av λ enbart gjordes utifrån BIC. Avagyan et al. redovisade heller inte glesheten hos estimaten i sin studie. Detta anser vi resulterar i att deras utvärderingsmått blir svåra att jämföra och att deras slutsats eventuellt kan behöva revideras.

5.2 Att välja λ för att jämföra estimat

Anledningen till att det är problematiskt att jämföra resultaten för olika k då λ väljs med BIC, är alltså att estimaten för olika k kan bli olika glesa. Det kan därför vara svårt att veta om det är valet av k eller glesheten i sig som påverkar resultaten. Det är till exempel naturligt att ett estimerat nätverk med få länkar också har färre noder som har möjlighet att interagera med andra noder och därmed få höga centralitetsmått. Det blir därför en systematisk diskriminering eftersom nätverkens storlek i hög grad påverkar utvärderingsmått. Detta medför att jämförelsen av utvärderingsmått vid olika k i dessa fall kommer handla mer om huruvida BIC fungerar tillfredsställande och inte så mycket om jämförelsen mellan olika k.

En mer rättvis jämförelse mellan olika värden på k erhålls alltså då estimatens gleshet eller FDR fixeras. Svårigheten kan då vara att fixera vid en gleshet som estimatet har möjlighet att prestera bra på. I Figur 14 i Avsnitt 4.3.5 såg vi att genom att välja λ med BIC blev nätverket glesare än för det estimat där glesheten fixerats. I båda figurerna ser det ut som att ungefär samma antal sanna länkar estimerats, men då estimatet som är mindre glest tvingas estimera fler länkar så blir de till större del falska. Det hade varit önskvärt att fixera glesheten vid några olika nivåer för att se om ett och samma k alltid presterar bäst. Det hade varit nödvändigt att fixera vid fler och framför allt högre nivåer av FDR, då det för vissa nätverksmodeller och k-värden aldrig gjordes estimat med lägre FDR än 0.5.

Det som däremot går att avläsa från resultaten då λ valts med BIC är att k-glasso med $k > 1$ i alla fall utom ett estimerar ett glesare nätverk än det som $k = 1$ ger. Eftersom de glesare nätverken får lägre FDR kan slutsatsen dras att BIC verkar fungera bättre för de högre rötterna om FDR är ett mått som är av stor vikt.

5.3 Problem vid generering av nätverk

Ökningen av styrkan på länkarna (signalstyrkan) på diagonalen för att få matriser inverterbara (se Avsnitt 3.1) medför en störning av den relativa signalstyrkan. Det är osäkert hur det har påverkat estimeringen och skulle behöva utredas ytterligare. Vidare kan det undersökas hur pass stor ökning av diagonalens magnitud som kan tillåtas utan att detta medför en allt för stor effekt på den relativa signalstyrkan i blocken.

5.4 Användning av likelihood-baserade mått

Det finns en avvägning som måste göras angående vilka utvärderingsmått som ska väga tungt beroende på estimeringssyftet. Om syftet är att hitta specifika länkar som kan ha betydelse för ett nätverk, är det viktigt att de länkar som estimerats i hög grad är sanna. Då är ett estimat med färre länkar men där de flesta av dem är sanna, att föredra framför ett estimat som lyckats fånga det sanna nätverkets struktur, till exempel blockstruktur, men kanske inte prickar in exakt rätt länkar. Om syftet istället är att estimera det sanna nätverkets fördelning och inte exakta länkar kan MSE vara ett bra mått. Vid de tillämpningar som den här rapporten har i åtanke kan syftet vara att hitta länkar som är aktiva vid exempelvis cancer, som inte tidigare är kända och skulle kunna öka kunskapen om vad som orsakar cancer.

Efter simuleringen upptäcktes att MSE, på grund av en illa vald avrundningsprecision, i många fall blev för liten för att jämföras. Eftersom simuleringarna var tidskrävande med så pass högdimensionella nätverksmodeller och många replikat hann vi inte åtgärda detta. Ett sätt att lösa detta på hade varit att skala upp MSE. De få synbara resultat vi fick tydde på att MSE visade på ett motsatt resultat jämfört med de övriga måtten. Detta kan bero på att MSE tar hänsyn till magnituden på de estimerade länkarna snarare än existensen av en länk, se Avsnitt 2.5.2. Eftersom vi var intresserade av förekomsten av länkar och inte magnituden av dem, anser vi att MSE inte är ett av de viktigaste utvärderingsmåtten för vårt syfte. Originalartikelns resultat lägger stor bevisbörda på just detta mått, vilket vi avstår från att göra i vår studie, av hänsyn till detta argument.

5.5 Placering av felestimerade länkar

I resultat presenterades en jämförelse över var i estimatet som majoriteten av de felestimerade länkarna placerades. Betydelsen av huruvida det finns en systematisk skillnad i var k-glasso och glasso tenderar att felestimera länkar beror på vad som är av intresse för den specifika datan. Om vi till exempel antar att blocken representerar en biologisk process och det är en process som vi inte är särskilt intresserade av, blir interaktioner mellan variabler i den processen inte intressanta. Däremot kan en interaktion utanför en sådan process vara ett samband som inte tidigare är känt och är då mer intressant. Då skulle en metod som i kombination med låg FDR och som heller inte estimerade så många felaktiga länkar utanför blocken, vara en fördel. Resultaten av simuleringen tydde på att k-glasso med $k = 2$ skulle kunna ha ett sådant mönster.

5.6 Prestation av k-glasso för olika nätverk

I vår studie gav roten $k = 2$ generellt bättre resultat än övriga rötter för nätverken Prec U200/U200+. Prec U200 var det nätverket som användes från originalartikelns studie. För övriga nätverk som vi simulerat gav både roten $k = 1$ och $k = 1.5$ i regel mer fördelaktiga resultat. Skillnaden mellan nätverken var dels att de förstnämnda genererade data med högre egenvärdesspridning och dels att de hade olika fördelning. Det tyder på att valet av k är dataspecifikt.

Eftersom grundidén med transformationen var att minska egenvärdesspridningen borde de nätverksmodeller med störst egenvärdesspridning gagnas mest av en egenvärdestransformation. Då den högre roten presterade bäst för just den datan som har större egenvärdesspridning verkar det antagandet stämma. Frågan är huruvida det finns en gräns för hur stor egenvärdesspridningen bör vara om $k = 2$ ska hjälpa snarare än stjälpa. Om egenvärdesspridningen inte är stor nog kanske transformationen orsakar att struktur i datan går förlorad snarare än att den dämpar det dominanta till fördel för att se detaljer.

6 Slutsats

I den här rapporten har vi undersökt en ny metod för att estimera stora glesa nätverk. Slutsatsen i artikeln av Avagyan et al. [1] var att $k = 2$ i praktiken är en generellt bra inställning på k. Detta kan vi inte bestyrka då våra resultat skiljer sig avsevärt mellan de olika nätverksmodellerna. Det verkar dock troligt att valet av k beror på hur pass stor egenvärdesspridning datan uppvisar. Förutom att valet av k är dataspecifikt föreslår vi att det även är fördelningsspecifikt och λ-specifikt. I vissa fall fann vi att k-glasso presterade sämre än glasso och det vore därför värdefullt att fastslå om det finns specifika förutsättningar då det är lämpligt respektive olämpligt att använda k-glasso framför glasso. För framtida studier föreslår vi att undersöka valet av k-rot baserat på egenvärdesspridning och att valet av λ undersöks med exempelvis korsvalidering och modellselektion.

Referenser

[1] Vahe Avagyan, Andrés M Alonso, and Francisco J Nogales. Improving the graphical lasso estimation for the precision matrix through roots of the sample covariance matrix. *Journal of Computational and Graphical Statistics*, pages 1–8, 2017.

[2] Peter Langfelder and Steve Horvath. Wgcna: an r package for weighted correlation network analysis. *BMC bioinformatics*, 9(1):559, 2008.

[3] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

[4] Kunihiro Baba, Ritei Shibata, and Masaaki Sibuya. Partial correlation and conditional correlation as measures of conditional independence. *Australian & New Zealand Journal of Statistics*, 46(4):657–664, 2004.

[5] S.L. Lauritzen. *Graphical models*. Oxford University Press, Oxford, 1996.

[6] Daniela M Witten, Jerome H Friedman, and Noah Simon. New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900, 2011.

[7] Teresia Kling, Patrik Johansson, José Sanchez, Voichita D Marinescu, Rebecka Jörnsten, and Sven Nelander. Efficient exploration of pan-cancer networks by generalized covariance selection and interactive web content. *Nucleic acids research*, 43(15):e98–e98, 2015.

[8] Onureena Banerjee, Laurent El Ghaoui, and Alexandre d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

[9] Iain M Johnstone. On the distribution of the largest eigenvalue in principal components analysis. *Annals of statistics*, pages 295–327, 2001.

[10] David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*, volume 571. John Wiley & Sons, 2005.

[11] Andrew Y Ng, Michael I Jordan, and Yair Weiss. On spectral clustering: Analysis and an algorithm. In *Advances in neural information processing systems*, pages 849–856, 2002.

[12] Ming Yuan and Yi Lin. Model selection and estimation in the gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.

[13] Rajul Parikh, Annie Mathai, Shefali Parikh, G Chandra Sekhar, and Ravi Thomas. Understanding and using sensitivity, specificity and predictive values. *Indian journal of ophthalmology*, 56(1):45, 2008.

[14] Karel GM Moons, Gerrit-Anne van Es, Jaap W Deckers, J Dik F Habbema, and Diederick E Grobbee. Limitations of sensitivity, specificity, likelihood ratio, and bayes' theorem in assessing diagnostic probabilities: a clinical example. *Epidemiology*, pages 12–17, 1997.

[15] John D Storey and Robert Tibshirani. Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, 100(16):9440–9445, 2003.

[16] Erik Holst and Poul Thyregod. A statistical test for the mean squared error. *Journal of Statistical Computation and Simulation*, 63(4):321–347, 1999.

[17] Das Kouisk, Samanta Sovan, and Pal Madhumangal. Study on centrality measures in social networks: a survey. *Social Network Analysis and Mining*, 2018.

[18] Eitan Gross. Statistical mechanics of scale-free gene expression networks. *EPL (Europhysics Letters)*, 100(5):58004, 2012.

[19] Zhiwei Wang, Sanjeev Banerjee, Dejuan Kong, Yiwei Li, and Fazlul H Sarkar. Down-regulation of forkhead box m1 transcription factor leads to the inhibition of invasion and angiogenesis of pancreatic cancer cells. *Cancer research*, 67(17):8293–8300, 2007.

Appendix

I tabeller över utvärderingsmått anger första raden för varje k medelvärdet av utvärderingsmåtten och nedre raden är standardavvikelser inom parentes. Tal är angivna med tre signifikanta siffror. Förkortningarna står för specificitet, sensitivitet, falsk upptäcktsfrekvens, medelkvadratfel, gleshet, gradcentralitet och sammankopplingscentralitet i angiven ordning.

Bilaga A Utvärderingsmått för nätverk valt med BIC

Tabell A1: *Utvärderingsmått för nätverk valt med BIC, Prec SF200 och Prec SF500.*

	SPEC	SENS	FDR	MSE	Gleshet	g-centr.	s-centr.
Prec SF200, n=80							
k = 1	0.993	0.386	0.383	0.020	0.983	0.358	0.378
	(0.001)	(0.014)	(0.043)	(0.001)	(0.001)	(0.059)	(0.070)
k = 1.5	0.998	0.354	0.192	0.023	0.988	0.337	0.347
	(0.001)	(0.013)	(0.034)	(0.001)	(0.001)	(0.048)	(0.063)
k = 2	0.997	0.259	0.311	0.013	0.990	0.143	0.100
	(0.002)	(0.017)	(0.100)	(0.000)	(0.002)	(0.070)	(0.000)
Prec SF500, n=200							
k = 1	0.994	0.382	0.447	0.038	0.986	0.406	0.470
	(0.000)	(0.010)	(0.014)	(0.001)	(0.001)	(0.051)	(0.032)
k = 1.5	0.998	0.326	0.212	0.089	0.992	0.395	0.478
	(0.000)	(0.009)	(0.019)	(0.001)	(0.000)	(0.060)	(0.034)
k = 2	0.993	0.151	0.680	0.027	0.991	0.130	0.100
	(0.001)	(0.006)	(0.020)	(0.000)	0.001)	(0.019)	(0.000)
Prec SF200, n=1000							
k = 1	0.985	0.775	0.407	0.005	0.965	0.618	0.592
	(0.002)	(0.011)	(0.024)	(0.000)	(0.002)	(0.052)	(0.061)
k = 1.5	0.994	0.739	0.219	0.000	0.975	0.611	0.651
	(0.001)	(0.012)	(0.026)	(0.054)	(0.001)	(0.052)	(0.051)
k = 2	0.998	0.238	0.187	0.001	0.992	0.110	0.100
	(0.000)	(0.005)	(0.032)	(0.050)	(0.000)	(0.000)	(0.000)
Prec SF500, n=2500							
k = 1	0.989	0.722	0.427	0.008	0.975	0.578	0.515
	(0.000)	(0.007)	(0.009)	(0.000)	(0.001)	(0.045)	(0.031)
k = 1.5	0.996	0.659	0.215	0.103	0.983	0.573	0.546
	(0.000)	(0.005)	(0.009)	(0.000)	(0.000)	(0.038)	(0.026)
k = 2	0.997	0.126	0.532	0.109	0.995	0.100	0.100
	(0.000)	(0.001)	(0.030)	(0.000)	(0.000)	(0.000)	(0.000)

Tabell A2: *Utvärderingsmått för nätverk valt med BIC, Prec U200 och Prec U200+.*

	SPEC	SENS	FDR	MSE	Gleshet	g-centr.	s-centr.
Prec U200, n=80							
k = 1	0.928	0.182	0.740	0.000	0.915	0.152	0.133
	(0.003)	(0.007)	(0.008)	(0.000)	(0.003)	(0.045)	(0.056)
k = 1.5	0.940	0.176	0.714	0.000	0.926	0.135	0.135
	(0.002)	(0.006)	(0.008)	(0.000)	(0.002)	(0.042)	(0.054)
k = 2	0.951	0.162	0.684	0.000	0.937	0.136	0.123
	(0.006)	(0.012)	(0.023)	(0.000)	(0.006)	(0.051)	(0.057)
Prec U200+, n=80							
k = 1	0.956	0.108	0.743	0.000	0.948	0.117	0.110
	(0.002)	(0.005)	(0.011)	(0.000)	(0.002)	(0.050)	(0.058)
k = 1.5	0.969	0.096	0.696	0.000	0.961	0.113	0.115
	(0.002)	(0.004)	(0.012)	(0.000)	(0.002)	(0.050)	(0.055)
k = 2	0.963	0.103	0.716	0.001	0.955	0.069	0.105
	(0.005)	(0.010)	(0.025)	(0.000)	(0.005)	(0.052)	(0.052)
Prec U200, n=1000							
k = 1	0.936	0.320	0.590	0.000	0.905	0.157	0.199
	(0.005)	(0.009)	(0.013)	(0.000)	(0.005)	(0.037)	(0.039))
k = 1.5	0.952	0.335	0.508	0.000	0.917	0.183	0.203
	(0.001)	(0.006)	(0.008)	(0.000)	(0.001)	(0.048)	(0.044)
k = 2	0.978	0.474	0.246	0.000	0.924	0.359	0.174
	(0.004)	(0.017)	(0.032)	(0.000)	(0.005)	(0.072)	(0.070)
Prec U200+, n=1000							
k = 1	0.968	0.200	0.533	0.000	0.947	0.149	0.141
	(0.001)	(0.004)	(0.011)	(0.000)	(0.001)	(0.045)	(0.054)
k = 1.5	0.980	0.196	0.418	0.000	0.958	0.127	0.137
	(0.001)	(0.004)	(0.013)	(0.000)	(0.001)	(0.043)	(0.058)
k = 2	0.975	0.297	0.370	0.000	0.942	0.045	0.146
	(0.004)	(0.013)	(0.036)	(0.000)	(0.004)	(0.045)	(0.070)

Bilaga B Utvärderingsmått för nätverk valt utifrån gleshet

Tabell B1: *Utvärderingsmått för nätverk valt utifrån gleshet, Prec SF200 och Prec SF500.*

	SPEC	SENS	FDR	MSE	Gleshet	g-centr.	s-centr.
Prec SF200, n=80							
k = 1	0.985	0.438	0.560	0.017	0.974	0.417	0.393
	(0.001)	(0.012)	(0.014)	(0.001)	(0.001)	(0.064)	(0.067)
k = 1.5	0.985	0.435	0.559	0.019	0.974	0.411	0.380
	(0.001)	(0.013)	(0.019)	(0.000)	(0.001)	(0.068)	(0.067)
k = 2	0.982	0.353	0.648	0.028	0.973	0.251	0.138
	(0.001)	(0.008)	(0.009)	(0.005)	(0.001)	(0.034)	(0.049)
Prec SF500, n=200							
k = 1	0.989	0.434	0.566	0.031	0.980	0.466	0.460
	(0.000)	(0.007)	(0.007)	(0.001)	(0.000)	(0.041)	(0.030)
k = 1.5	0.989	0.420	0.575	0.078	0.980	0.461	0.456
	(0.000)	(0.006)	(0.008)	(0.000)	(0.000)	(0.040)	(0.032)
k = 2	0.984	0.211	0.790	0.037	0.980	0.102	0.100
	(0.000)	(0.009)	(0.008)	(0.004)	(0.000)	(0.024)	(0.000)
Prec SF200, n=1000							
k = 1	0.993	0.726	0.273	0.006	0.973	0.628	0.636
	(0.000)	(0.009)	(0.013)	(0.000)	(0.001)	(0.054)	(0.048)
k = 1.5	0.993	0.748	0.242	0.024	0.974	0.603	0.653
	(0.001)	(0.009)	(0.017)	(0.000)	(0.001)	(0.053)	(0.051)
k = 2	0.982	0.346	0.655	0.021	0.973	0.239	0.104
	(0.001)	(0.004)	(0.008)	(0.001)	(0.001)	(0.023)	(0.015)
Prec SF500, n=2500							
k = 1	0.993	0.668	0.335	0.011	0.980	0.573	0.524
	(0.000)	(0.007)	(0.007)	(0.000)	(0.000)	(0.055)	(0.034)
k = 1.5	0.994	0.695	0.299	0.102	0.980	0.565	0.535
	(0.000)	(0.004)	(0.007)	(0.000)	(0.000)	(0.041)	(0.031)
k = 2	0.984	0.210	0.790	0.065	0.980	0.107	0.100
	(0.000)	(0.004)	(0.005)	(0.001)	(0.001)	(0.010)	(0.000)

Tabell B2: *Utvärderingsmått för nätverk valt utifrån gleshet, Prec U200 och Prec U200+.*

	SPEC	SENS	FDR	MSE	Gleshet	g-centr.	s-centr.
Prec U200, n=80							
k = 1	0.893	0.227	0.773	0.000	0.879	0.149	0.130
	(0.001)	(0.007)	(0.007)	(0.000)	(0.000)	(0.042)	(0.059)
k = 1.5	0.894	0.233	0.767	0.000	0.879	0.138	0.127
	(0.001)	(0.006)	(0.006)	(0.000)	(0.001)	(0.041)	(0.050)
k = 2	0.894	0.234	0.766	0.000	0.879	0.141	0.117
	(0.002)	(0.011)	(0.011)	(0.000)	(0.001)	(0.057)	(0.062)
Prec U200+, n=80							
k = 1	0.886	0.190	0.810	0.000	0.876	0.107	0.108
	(0.001)	(0.007)	(0.007)	(0.000)	(0.001)	(0.052)	(0.058)
k = 1.5	0.886	0.193	0.807	0.001	0.876	0.107	0.108
	(0.002)	(0.006)	(0.006)	(0.000)	(0.001)	(0.053)	(0.056)
k = 2	0.885	0.185	0.815	0.001	0.876	0.097	0.111
	(0.002)	(0.009)	(0.009)	(0.000)	(0.002)	(0.054)	(0.060)
Prec U200, n=1000							
k = 1	0.911	0.358	0.642	0.000	0.879	0.155	0.191
	(0.001)	(0.006)	(0.006)	(0.000)	(0.001)	(0.036)	(0.038)
k = 1.5	0.916	0.389	0.611	0.000	0.879	0.174	0.196
	(0.001)	(0.006)	(0.006)	(0.000)	(0.001)	(0.047)	(0.044)
k = 2	0.940	0.566	0.434	0.000	0.879	0.266	0.164
	(0.002)	(0.014)	(0.014)	(0.000)	(0.001)	(0.062)	(0.055)
Prec U200+, n=1000							
k = 1	0.901	0.301	0.699	0.000	0.876	0.143	0.128
	(0.001)	(0.007)	(0.007)	(0.000)	(0.001)	(0.048)	(0.054)
k = 1.5	0.903	0.312	0.688	0.000	0.877	0.144	0.125
	(0.002)	(0.007)	(0.007)	(0.000)	(0.001)	(0.051)	(0.061)
k = 2	0.917	0.414	0.586	0.000	0.877	0.077	0.124
	(0.002)	(0.014)	(0.014)	(0.000)	(0.001)	(0.051)	(0.064)

Bilaga C Falskt positiva länkar

Notera att måttet FP kan vara missvisande om vi inte tar hänsyn till totala antalet positiva länkar.

Tabell C1: *Estimatens placering av FP (falskt positiva) länkar, angivet i andelar, för Prec SF200 och Prec SF500. Estimat utvalt med BIC.*

	Inom block	Utanför block
Prec SF200, n=80		
k = 1	0.297	0.703
	(0.046)	(0.046)
k = 1.5	0.463	0.537
	(0.086)	(0.086)
k = 2	0.999	0.001
	(0.006)	(0.006)
Prec SF500, n=200		
k = 1	0.193	0.807
	(0.011)	(0.011)
k = 1.5	0.356	0.644
	(0.033)	(0.033)
k = 2	0.907	0.093
	(0.043)	(0.043)
Prec SF200, n=1000		
k = 1	0.375	0.625
	(0.027)	(0.027)
k = 1.5	0.622	0.378
	(0.045)	(0.045)
k = 2	1.000	0.000
	(0.000)	(0.000)
Prec SF500, n=2500		
k = 1	0.267	0.733
	(0.011)	(0.011)
k = 1.5	0.510	0.490
	(0.021)	(0.021)
k = 2	0.931	0.069
	(0.004)	(0.004)

Tabell C2: *Estimatens placering av FP (falskt positiva) länkar, angivet i andelar, för Prec U200 och Prec U200+. Estimat utvalt med BIC.*

	Inom block	Utanför block
Prec U200, n=80		
k = 1	0.142	0.858
	(0.009)	(0.009)
k = 1.5	0.144	0.856
	(0.011)	(0.011)
k = 2	0.147	0.853
	(0.017)	(0.017)
Prec U200+, n=80		
k = 1	0.150	0.850
	(0.011)	(0.011)
k = 1.5	0.151	0.849
	(0.013)	(0.013)
k = 2	0.154	0.846
	(0.021)	(0.021)
Prec U200, n=1000		
k = 1	0.141	0.859
	(0.010)	(0.010)
k = 1.5	0.142	0.858
	(0.011)	(0.011)
k = 2	0.445	0.555
	(0.068)	(0.068)
Prec U200+, n=1000		
k = 1	0.151	0.849
	(0.013)	(0.013)
k = 1.5	0.163	0.837
	(0.019)	(0.019)
k = 2	0.275	0.725
	(0.043)	(0.043)

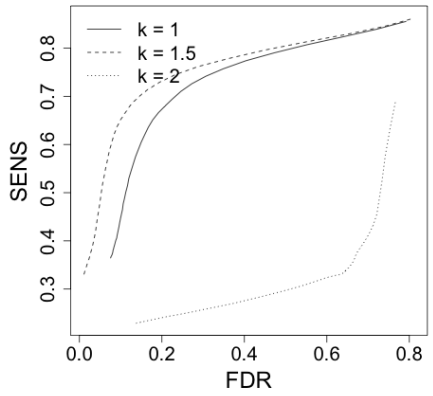
Tabell C3: *Estimatens placering av FP (falskt positiva) länkar, angivet i andelar, för Prec SF200 och Prec SF500. Estimat utvalt utifrån gleshet.*

	Inom block	Utanför block
Prec SF200, n=80		
k = 1	0.279	0.721
	(0.028)	(0.028)
k = 1.5	0.307	0.693
	(0.029)	(0.029)
k = 2	0.936	0.064
	(0.054)	(0.054)
Prec SF500, n=200		
k = 1	0.217	0.783
	(0.009)	(0.009)
k = 1.5	0.291	0.709
	(0.012)	(0.012)
k = 2	0.886	0.114
	(0.033)	(0.033)
Prec SF200, n=1000		
k = 1	0.422	0.578
	(0.040)	(0.040)
k = 1.5	0.601	0.399
	(0.046)	(0.046)
k = 2	0.989	0.011
	(0.005)	(0.005)
Prec SF500, n=2500		
k = 1	0.248	0.752
	(0.012)	(0.012)
k = 1.5	0.487	0.513
	(0.018)	(0.018)
k = 2	0.857	0.143
	(0.015)	(0.015)

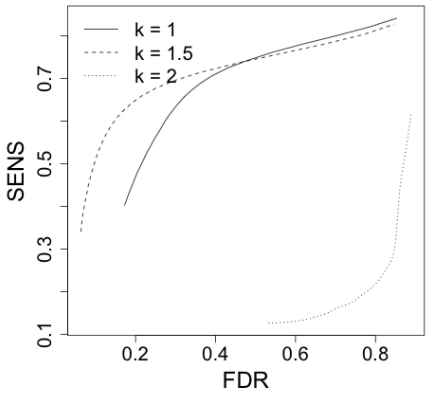
Tabell C4: *Estimatens placering av FP (falskt positiva) länkar, angivet i andelar, för Prec U200 och Prec U200+. Estimat utvalt utifrån gleshet.*

	Inom block	Utanför block
Prec U200, n=80		
k = 1	0.142	0.858
	(0.008)	(0.008)
k = 1.5	0.143	0.857
	(0.008)	(0.008)
k = 2	0.143	0.857
	(0.010)	(0.010)
Prec U200+, n=80		
k = 1	0.147	0.853
	(0.007)	(0.007)
k = 1.5	0.147	0.853
	(0.007)	(0.007)
k = 2	0.147	0.853
	(0.010)	(0.010)
Prec U200, n=1000		
k = 1	0.142	0.858
	(0.008)	(0.008)
k = 1.5	0.143	0.857
	(0.008)	(0.008)
k = 2	0.277	0.723
	(0.025)	(0.025)
Prec U200+, n=1000		
k = 1	0.147	0.853
	(0.008)	(0.008)
k = 1.5	0.147	0.853
	(0.008)	(0.008)
k = 2	0.190	0.810
	(0.017)	(0.017)

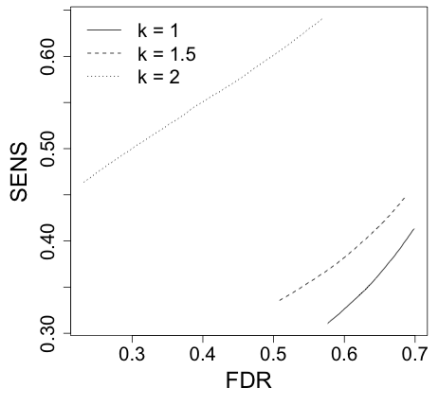
Bilaga D FDR och sensitivitet



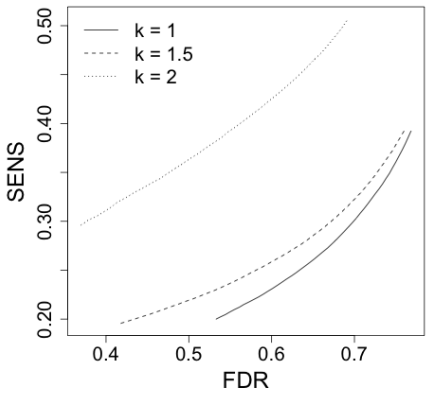
(a) Prec SF200



(b) Prec SF500

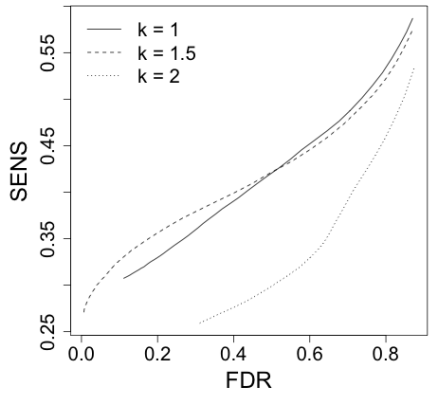


(c) Prec U200

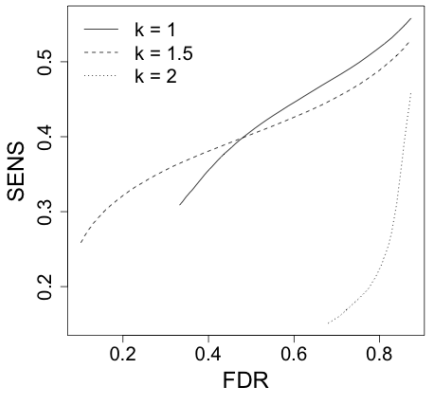


(d) Prec U200+

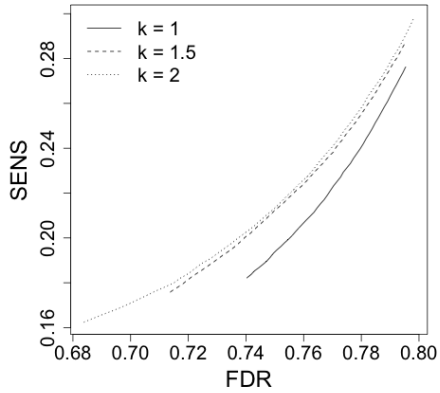
Figur D1: *Sensitivitet mot FDR som en funktion av en avtagande sekvens av* λ*. Eftersom det är önskvärt med hög sensitivitet och låg FDR är det fördelaktigt med en snabbt växande konkav kurva, som* $k = 1.5$ *i (a). En längre kurva betyder en större förändring i måtten över* λ*-sekvenserna. Dessa figurer gäller för nätverk när* $p < n$*.*



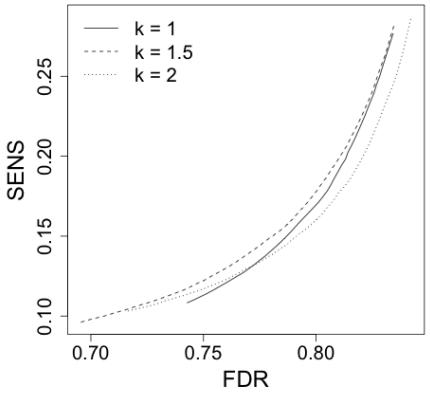
(a) Prec SF200



(b) Prec SF500



(c) Prec U200



(d) Prec U200+

Figur D2: *Sensitivitet mot FDR som en funktion av en avtagande sekvens av* λ*. Eftersom det är önskvärt med hög sensitivitet och låg FDR är det fördelaktigt med en snabbt växande konkav kurva. En längre kurva betyder en större förändring i måtten över* λ*-sekvenserna. Dessa figurer gäller för nätverk när* $p > n$*.*

Bilaga E Kod

https://github.com/hansronald/kandidaten