Right singular vector projection graphs: fast high dimensional covariance matrix estimation under latent confounding

Originally published in:
Journal of the Royal Statistical Society Series B: Statistical Methodology 82(2), https://doi.org/10.1111/rssb.12359

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2020)
82, *Part* 2, *pp.* 361–389

Right singular vector projection graphs: fast high dimensional covariance matrix estimation under latent confounding

Rajen D. Shah

University of Cambridge, UK

and Benjamin Frot, Gian-Andrea Thanei and Nicolai Meinshausen

Eidgenössiche Technische Hochschule Zürich, Switzerland

[Received January 2019. Revised November 2019]

Summary. We consider the problem of estimating a high dimensional $p \times p$ covariance matrix Σ, given n observations of confounded data with covariance $\Sigma + \Gamma\Gamma^{\mathsf{T}}$, where Γ is an unknown $p \times q$ matrix of latent factor loadings. We propose a simple and scalable estimator based on the projection onto the right singular vectors of the observed data matrix, which we call right singular vector projection (RSVP). Our theoretical analysis of this method reveals that, in contrast with approaches based on the removal of principal components, RSVP can cope well with settings where the smallest eigenvalue of $\Gamma^{\mathsf{T}}\Gamma$ is relatively close to the largest eigenvalue of Σ, as well as when the eigenvalues of $\Gamma^{\mathsf{T}}\Gamma$ are diverging fast. RSVP does not require knowledge or estimation of the number of latent factors q, but it recovers Σ only up to an unknown positive scale factor. We argue that this suffices in many applications, e.g. if an estimate of the correlation matrix is desired. We also show that, by using subsampling, we can further improve the performance of the method. We demonstrate the favourable performance of RSVP through simulation experiments and an analysis of gene expression data sets collated by the GTEX consortium.

Keywords: Causal structure learning; Covariance matrix; Graphical models; High dimensional data; Latent confounding

1. Introduction

Suppose that a random vector $w \in \mathbb{R}^p$ follows a multivariate normal distribution with covariance matrix Σ:

$$w \sim \mathcal{N}_p(\mu_w, \Sigma).$$

Given n independent and identically distributed (IID) copies of w whose rows form a data matrix $W \in \mathbb{R}^{n \times p}$ it is often of interest to estimate either Σ, or certain quantities that are derived from this such as the precision matrix $\Omega := \Sigma^{-1}$ or collections of conditional independences that may then be used to infer causal structure (Spirtes *et al.*, 2000).

Suppose now that we cannot observe W directly, but we instead observe n IID copies of a random vector x which form the rows of $X \in \mathbb{R}^{n \times p}$; x is related to w through

$$x = w + \Gamma h. \tag{1}$$

Address for correspondence: Rajen D. Shah, Statistical Laboratory, Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge, CB3 0WB, UK.
E-mail: R.Shah@statslab.cam.ac.uk

Here $h \in \mathbb{R}^q$ is a vector of unobserved latent random variables, and $\Gamma \in \mathbb{R}^{p \times q}$ a fixed matrix of loadings. If we assume that h is normally distributed, without loss of generality we may take $h \sim \mathcal{N}_q(\mu_h, I)$. We then have that the covariance Θ of the observed x contains a contribution $\Gamma\Gamma^{\mathrm{T}}$ from latent confounding and a contribution Σ from idiosyncratic noise:

$$\Theta = \mathrm{cov}(x) = \Gamma\Gamma^{\mathrm{T}} + \Sigma.$$

If we simply ignore the confounding, we shall have the covariance Θ as the target of inference instead of Σ, and the two can be very different.

Applications where such confounding is important in practice include the following.

(a) Cell biology: the activities of proteins and messenger ribonucleic acid, for example, can be confounded by environmental factors. Two highly correlated protein activities are thus not necessarily close in a causal network (Leek and Storey, 2007; Stegle *et al.*, 2012).
(b) Financial assets: the returns of various stocks will be confounded by some latent factors (such as general market movement or sector influences) without the covariance necessarily revealing anything about causal connections between companies (Menchero *et al.*, 2010).
(c) Confounding in biology and genetics can also occur due to technical malfunction and laboratory effects (Gagnon-Bartsch *et al.*, 2013).

Thus, in various settings, to infer meaningful connections between variables we would like to remove the effect of confounding from the empirical covariance $\hat{\Theta}$ of X and to estimate Σ.

As well as the intrinsic ill-posedness of the problem of separating Σ from a noisy observation of $\Sigma + \Gamma\Gamma^{\mathrm{T}}$ with Γ unknown, a further challenge in the applications above and many others is that the dimension p may be very large indeed; of the order of thousands or more. This high dimensionality brings computational difficulties that must be addressed by any practical procedure.

For Σ to be identifiable, appropriate assumptions on both Σ and Γ must be made. One natural assumption is that the minimum eigenvalue γ_l of $\Gamma^{\mathrm{T}}\Gamma$ is larger than the largest eigenvalue σ_u of Σ. In this setting, a popular strategy to deal with unwanted confounding is removal of top principal components from $\hat{\Theta}$. This has been proposed in Gagnon-Bartsch *et al.* (2013) and Fan *et al.* (2013). The latter work, a discussion paper in the *Journal of the Royal Statistical Society*, Series B, shows that, when σ_u is bounded and $\gamma_l = O(p)$, and so the gap between the quantities is large, Σ may be recovered consistently. In this case the top q eigenvalues of $\hat{\Theta}$ will be well separated from the rest, and so exactly q principal components can be removed from $\hat{\Theta}$: this is important as removing too many or too few principal components can result in a poor estimate.

However, as several discussants of Fan *et al.* (2013) pointed out, in many settings empirical covariances do not display well-separated eigenvalues even when latent factors are known to be present. When the gap between σ_u and γ_l is not sufficiently large, the top q eigenvalues can be close to the bulk, making estimation of q challenging and potentially impossible (Barigozzi and Cho, 2018). Furthermore the top principal components (PCs) of the empirical covariance can be far from those of Θ (Donoho *et al.*, 2018), so, even if q were known, the PC removal approach would not work well.

In this paper, we propose a simple approach to estimating Σ that can cope with settings where the gap between γ_l and σ_u may range from large and $O(p)$ to potentially small. To achieve this ambitious objective, the method sacrifices estimation of the scale of Σ: we recover Σ only up to an unknown positive scalar factor. The loss of scale, however, is inconsequential when the ultimate goal is rather to estimate the correlation matrix $\tilde{\Sigma}$, or to locate the top s largest entries in Σ for a prespecified s, to build a network. In fact, we show that the scale-free nature of our estimator gives it an in-built robustness in that, if the rows of X have elliptical distributions, its distribution is precisely the same as if the data were Gaussian (see proposition 3 in Section 3).

Let $V \in \mathbb{R}^{p\times(n-1)}$ be the matrix of right singular vectors with non-zero singular values of a column-centred version of X. Our estimator is based on $\hat{\Sigma}_{\text{rsvp}} := VV^{\text{T}}$; we call this right singular vector projection (RSVP). The PC removal estimate is proportional to VH^2V^{T} where H is a diagonal matrix of singular values of the centred X with the first q entries set to 0 (when q is known). Thus RSVP may be seen as a highly regularized version of PC removal, where the random H is set to the identity matrix to reduce its variance. In fact, we show that each entry of $\hat{\Sigma}_{\text{rsvp}}$ concentrates around its expectation at the same rate as the empirical covariance matrix after rescaling, even in settings where q is allowed to grow at almost the same rate as n (see theorem 1 in Section 3).

Despite the aggressive regularization, it turns out that the bias is dominated by the variance provided that $p \gg n$, so n/p is small. As a consequence, we can show that, with high probability,

$$\inf_{\kappa>0} \max_{j,k} |\Sigma_{jk} - \kappa\hat{\Sigma}_{\text{rsvp},jk}| \leqslant c\sqrt{\left\{\frac{\log(p)}{n}\right\}}$$

for some constant $c > 0$, even in certain settings when γ_l is only larger than σ_u by a constant factor, and the latter is bounded. In fact, we show that the statistical properties of $\hat{\Sigma}_{\text{rsvp}}$ are such that, when used as input to several standard procedures for conditional independence graph estimation or causal discovery procedures, the performances of the resulting estimates are, in many settings, identical to those attained when working with the unconfounded data, up to constant factors.

One requirement for $\hat{\Sigma}_{\text{rsvp}}$ to work well is that $p \gg n$. For settings where n is large, we can circumvent this condition by using a subsampling strategy. We show that, surprisingly, by computing our estimator on subsamples of the data and averaging (Breiman, 1996), the bias may be reduced, and the variance inflated by only a factor of $\sqrt{\{\log(p)\}}$. Subsampling with a very small number of samples in each subsample is both statistically and computationally attractive and is the approach that we would recommend in settings where we do not have $p \gg n$.

1.1. Related work

There is a large body of work on high dimensional covariance and precision matrix estimation: see for example the recent review paper of Cai *et al.* (2016) and references therein. Much of the work on the specific setting with latent confounding has focused on estimation of the precision matrix Ω, which is assumed to be sparse. The presence of the latent confounding causes the overall precision matrix of x to be a sum of Ω and a low rank component. One approach to sparse precision matrix estimation in the absence of confounding is the graphical lasso (Yuan and Lin, 2007; Yuan, 2010; Friedman *et al.*, 2008). Building on this and work on sparse–dense matrix decompositions in the noiseless setting (Candès *et al.*, 2011; Chandrasekaran *et al.*, 2011), the work of Chandrasekaran *et al.* (2012) formulates a convex objective involving nuclear norm penalization for Gaussian graphical model estimation with latent confounders. The work of Frot *et al.* (2019b) uses this as a stepping-stone for causal structure learning and causal effect estimation in low dimensional settings. A challenge for nuclear norm penalization and related approaches is that, although the objective is convex, optimizing it is nevertheless a computationally intensive task that does not scale to very large dimensions.

A second approach to precision matrix estimation exploits the fact that coefficients from regressions of each variable on all others, known as nodewise regressions, match the entries of the precision matrix up to scale factors (Lauritzen, 1996; Meinshausen and Bühlmann, 2006). Adjusting for confounding can be built into a nodewise regression procedure, e.g. by using the 'Lava' method of Chernozhukov *et al.* (2017) which employs a sparse–dense decomposition of

the regression coefficients; the sparse part of the coefficients can then be retained as the dense part is generally due to confounding. This regression may be formulated as a lasso regression with a transformed response and a particular preconditioned design matrix; see also Rohe (2015) for an earlier equivalent proposal. Ćevid *et al.* (2018) studied the theoretical properties of the Lava approach as well as more general forms of preconditioning including the Puffer transform that was proposed in Jia and Rohe (2015) and further investigated in Wang and Leng (2015). This, in analogy with RSVP, modifies the design matrix by replacing non-zero singular values with a constant. We also note that the asymptotic normal thresholding procedure of Ren *et al.* (2015), which employs nodewise regressions in a different fashion, is robust to weak confounding.

There has been comparatively less work on covariance matrix estimation in the presence of confounding, though, as we discuss and make use of in this work, an estimated covariance can be used as a starting point for conditional independence graph estimation or causal discovery. In addition to the work of Fan *et al.* (2013) and Gagnon-Bartsch *et al.* (2013) that was mentioned earlier, Fan *et al.* (2018) proposed a PC removal approach that can be applied to heavy-tailed data that follow an elliptical distribution.

1.2. Organization of the paper

The rest of the paper is organized as follows. In Section 2 we first discuss asymptotic identifiability of Σ and then introduce our RSVP estimator $\hat{\Sigma}_{\mathrm{rsvp}}$ and versions involving subsampling. We present theoretical properties of $\hat{\Sigma}_{\mathrm{rsvp}}$ and RSVP with sample splitting in Section 3. In Section 4 we present results on the use of RSVP estimators as input to methods for conditional independence graph estimation and for causal discovery via the PC algorithm. Numerical experiments are contained in Section 5 and we conclude with a discussion in Section 6. The on-line supplementary material for this paper contains all proofs and further results concerning the GTEX consortium data analyses that are presented in Section 5.2.

1.3. Notation

We write $a \lesssim b$ as shorthand for 'there is a constant $c > 0$ such that $a \leqslant cb$'. This constant may be a universal constant, or a function of quantities that have been designated as constants in our assumptions. If $a \lesssim b$ and $b \lesssim a$, we may write $a \asymp b$. For a matrix $A \in \mathbb{R}^{d \times m}$, $\|A\|$ will denote the operator norm, and $\|A\|_\infty = \max_{i=1,\ldots,d,\, j=1,\ldots,m} |A_{ij}|$.

When $d = m$, so A is square, we shall write $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ for the maximum and minimum eigenvalues of A respectively. Further, given sets $I, J \subseteq D := \{1, \ldots, d\}$, we shall denote by $A_{I,J}$ the $|I| \times |J|$ submatrix of A that is formed from those rows and columns of A indexed by I and J respectively. Such matrix subsetting operations will always be considered to have been performed first so that for example, when $A_{I,I}$ is invertible, $A_{I,I}^{-1} \equiv (A_{I,I})^{-1}$. For $j \in D$, j or $-j$ used in place of the subscripts I or J above will represent $\{j\}$ and $D \setminus \{j\}$ respectively, so $A_{j,-j}$ for example is the $1 \times (d-1)$ matrix that is formed from the jth row of A with its jth entry removed.

In analogy with the matrix subsetting notation that was set out above, we shall write, for a vector $v \in \mathbb{R}^d$, v_I for the subvector formed from the components of v indexed by I. Also, for $j, k \in D$, v_{-j} and v_{-jk} will be subvectors of v with jth and both the jth and kth components removed respectively. We denote by e_j the jth standard basis vector; the dimension of this will be clear from the context.

2. Right singular vector projection

Let us assume that the observed data matrix $X \in \mathbb{R}^{(n+1) \times p}$ has rows given by $n+1$ independent

realizations of an $\mathcal{N}_p(\mu, \Theta)$ random vector (we shall later relax the Gaussian assumption; see proposition 3 in Section 3). The $n+1$ rather than n is for mathematical convenience: the column-centred version $\tilde{X} := \Pi X$ of X effectively contains n observations. Here $\Pi = I - (n+1)^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}$ where $\mathbf{1}$ is an $(n+1)$-vector of 1s. Our goal is to construct an estimate of Σ based on these data where $\Sigma + \Gamma\Gamma^{\mathrm{T}} = \Theta$ and both $\Gamma \in \mathbb{R}^{p\times q}$ and q are unknown. We are interested in the case $p \gg n$ and shall assume that $p > cn$ for some $c > 1$, unless specified otherwise.

In what follows we first study the identifiability of Σ in the model above. In Section 2.2 we discuss a general approach for estimating Σ based on transforming the spectrum of the covariance matrix, which includes PC removal and our RSVP method presented in Section 2.3 as special cases. Finally we introduce a sample splitting version of RSVP in Section 2.4.

2.1. Asymptotic identifiability

First consider an artificial setting where Θ itself is directly observed. Even in this noiseless setting, certain conditions must be placed on Γ and Σ for Σ to be recoverable given Θ. Define

$$\begin{aligned}\lambda_{\min}(\Gamma^{\mathrm{T}}\Gamma) &:= \gamma_l,\\ \lambda_{\max}(\Gamma^{\mathrm{T}}\Gamma) &:= \gamma_u,\\ \lambda_{\min}(\Sigma) &:= \sigma_l,\\ \lambda_{\max}(\Sigma) &:= \sigma_u.\end{aligned}$$

If γ_l is large compared with σ_u, we might hope that the top q eigenvectors of Θ will span most of the column space of Γ. Therefore removing these from Θ should yield a matrix that is close to Σ. Proposition 1 below, based in part on an application of the Davis–Kahan $\sin(\theta)$ theorem (Davis and Kahan, 1970), formalizes this intuition.

Let Θ have eigendecomposition PD^2P^{T} where the diagonal matrix D has $D_{11} \geqslant D_{22} \geqslant \ldots \geqslant D_{pp}$. Also define, for $l \in \{1, \ldots, p\}$, function H_l taking as argument a square matrix, and outputting a matrix of the same dimension, by

$$(H_l(E))_{jk} = \begin{cases} 0 & \text{if } j, k \leqslant l,\\ E_{jk} & \text{otherwise.}\end{cases}$$

Thus the top left-hand $l \times l$ submatrix of $H_l(E)$ is a matrix of 0s. Define $\Pi_\Gamma := \Gamma(\Gamma^{\mathrm{T}}\Gamma)^{-1}\Gamma^{\mathrm{T}}$,

$$\rho_1 := \|\Pi_\Gamma \Sigma\|$$

and

$$\rho_2 := \max_j \|\Pi_\Gamma e_j\|_2.$$

Proposition 1. Suppose that σ_l is bounded away from 0 and $\gamma_l > c\sigma_u$ for a constant $c > 1$. Then

$$\|PH_q(D^2)P^{\mathrm{T}} - \Sigma\|_\infty \lesssim \rho_1\rho_2 + \gamma_u\rho_1^2/\gamma_l^2. \tag{2}$$

In order for removal of q PCs to yield a matrix that is close to Σ at the population level, we require ρ_2 to be small; this essentially requires that the column space of Γ is not too closely aligned with any of the standard basis vectors. We always have the bound $\rho_1 \leqslant \sigma_u$. However, in the setting where Γ is entirely uninformative about Σ, we might expect that ρ_1 may be smaller. Specifically, if we imagine that nature has chosen the column space of Γ uniformly at random, we shall have with high probability that

$$\rho_1^2 \lesssim \frac{1}{p}[\mathrm{tr}(\Sigma^2) + \sqrt{\{q\,\mathrm{tr}(\Sigma^4)\}}],$$
$$\rho_2^2 \lesssim \frac{q}{p}\left(1 + \max\left[\frac{\log(p)}{q}, \sqrt{\left\{\frac{\log(p)}{q}\right\}}\right]\right). \tag{3}$$

See section H in the on-line supplementary material for a derivation. Asymptotic identifiability results related to proposition 1 are given in Fan *et al.* (2013, 2018) when σ_u and q are bounded, and both γ_l and γ_u are $O(p)$. In these settings it is straightforward to show that $\rho_2 \lesssim p^{-1/2}$, in which case the right-hand side of expression (2) may be replaced by $p^{-1/2}$.

2.2. Spectral transformations

We now return to the original noisy version of the problem. The empirical covariance matrix $\hat{\Theta} = \tilde{X}^{\mathrm{T}}\tilde{X}/n$ has expectation $\Theta = PD^2P^{\mathrm{T}}$, so we would ideally like to modify $\hat{\Theta}$ such that the eigenstructure of its expectation more closely resembles $PH_q(D^2)P$. Therefore consider the following family of estimators that involve transforming the spectrum of $\hat{\Theta}$.

Note that, as $\tilde{X}$ has been centred and $p > n$, the rank of $\tilde{X}$ is n. Let the singular value decomposition of $\tilde{X}$ be given by $\tilde{X} = U\Lambda V^{\mathrm{T}}$ where $\Lambda \in \mathbb{R}^{n\times n}$ is diagonal, and $U \in \mathbb{R}^{(n+1)\times n}$ and $V \in \mathbb{R}^{p\times n}$ each have orthonormal columns. Define

$$\hat{\Sigma}_H = \frac{1}{n} VH(\Lambda^2)V^{\mathrm{T}} \tag{4}$$

where function H here outputs $n \times n$ diagonal matrices. For such estimators, we have the following property.

Proposition 2. We have that the expectation $\mathbb{E}(\hat{\Sigma}_H) = PC_H^2P^{\mathrm{T}}$ where C_H is diagonal.

The fact that the eigenvectors of $\mathbb{E}(\hat{\Sigma}_H)$ coincide with those of Θ suggests that we should pick function H such that C_H^2 is close to $H_q(D^2)$. A natural choice is a simple PC-analysis-based adjustment (Fan *et al.*, 2013, 2018; Gagnon-Bartsch *et al.*, 2013) of the form

$$\hat{\Sigma}_{\mathrm{pca}}(l) := \hat{\Sigma}_{H_l} = n^{-1}VH_l(\Lambda^2)V^{\mathrm{T}}.$$

The resulting PC removal estimator can be further thresholded as in Bickel and Levina (2008) and Fan *et al.* (2013), though, if our aim is to recover the locations of the largest entries of the covariance, this additional thresholding step is without consequence. The choice of the number l of principal components to remove is rather critical to the method but can be challenging. Even if we had knowledge about the dimensionality q of the latent confounders, the optimal choice would depend on the relative magnitude of the eigenvalues of $\Gamma^{\mathrm{T}}\Gamma$ in relation to the eigenvalues of Σ. In the absence of this knowledge, one might resort to cross-validation schemes. Since the target of inference is the unobserved idiosyncratic part Σ of the covariance, it is not obvious how such a cross-validation can be set up in a meaningful way. Information criteria may be used as in Fan *et al.* (2013), but these rely on $\gamma_l/\sigma_u \gtrsim p$.

2.3. Right singular vector projection

One reason that the PC removal approach can struggle in settings where the separation between γ_l and γ_u is relatively small is that the top q eigenvectors of $\hat{\Theta}$ need not span the column space of Γ well and in general will have high variability. Thus although $\hat{\Theta} = n^{-1}V\Lambda^2V^{\mathrm{T}}$ concentrates well around its expectation Θ in l_∞-norm, an approach that involves manipulating the contributions

of individual singular vectors in V to the overall estimator is likely to have high variance. This suggests that some form of regularization may be helpful.

Taking the function H as one which always returns n times the identity matrix results in the simple estimator

$$\hat{\Sigma}_{\text{rsvp}} := VV^{\mathrm{T}}.$$

Note that this is invariant to permutations of the columns of V and so is less dependent on properties of individual eigenvectors. As a consequence of the regularization, we have lost the scaling of the original covariance: the estimator is invariant to multiplying X from the left by any invertible $n \times n$ matrix. Thus we can only hope to recover Σ up to a constant scale factor. This suffices for our purposes, and we argue that this gives the estimator a certain robustness in that it is insensitive to particular pretransformations of the data such as scaling of the rows of X. In fact $\hat{\Sigma}_{\text{rsvp}}$ is more generally robust; see proposition 3 in Section 3. The computation time is dominated by the matrix multiplication of V and V^{T}, which is $O(np^2)$; thus the computational complexity is the same as that for computing the empirical covariance.

In a regression context, an analogous approach for preconditioning the design matrix has been explored in Jia and Rohe (2015) and Wang and Leng (2015). The Lava estimator (Chernozhukov *et al.*, 2017) employs a similar preconditioning strategy but, instead of setting all non-zero singular values of the design matrix to 1, the singular values d_i are transformed implicitly as $\{d_i^2/(1+cd_i^2)\}^{1/2}$, where the constant c depends on a tuning parameter and the sample size.

It may seem as if all information regarding the eigenvalues of Σ has been lost in the regularization as Λ does not play a role in the estimator. However, we show in Section 3 that, in certain high dimensional settings, we can even estimate Σ in l_∞-norm at the same rate as the empirical covariance matrix in the absence of confounding, though only up to an unknown scale factor. Intuitively, the reason is that, when $p \gg n$, with the exception of certain large eigenvalues in Λ due to large eigenvalues in $\Gamma^{\mathrm{T}}\Gamma$, the rest of the eigenvalues are essentially noise and bear no resemblance to the eigenvalues of Σ. This peculiar blessing of high dimensionality is a phenomenon that fails when p is of the same order as n, for example. It is, however, possible to subsample the data, and to average over estimates computed on the samples, to mimic the high dimensional setting. We discuss this below.

2.4. Subsampling right singular vector projection

Given $m \in \{1, \ldots, n\}$, let $V^{(b)}$ be the matrix of right singular vectors of a random sample of m rows of $\tilde{X}$. We define the subsampling RSVP estimator as

$$\hat{\Sigma}_{\text{rsvp-sub}} := \frac{1}{B}\sum_{b=1}^{B} V^{(b)}(V^{(b)})^{\mathrm{T}}.$$

The sample splitting RSVP estimator $\hat{\Sigma}_{\text{rsvp-split}}$ is defined similarly but where the sets of indices of the sampled rows are disjoint, and so $B = \lceil (n+1)/m \rceil$. In practice, the subsampling estimator is preferable as the additional sampling can help to reduce the variance of the estimator. Our main reason for introducing the sample splitting version is that it is simpler to understand its theoretical properties (see theorem 4 in Section 3); however, sample splitting still performs well empirically as we demonstrate in Section 5.

Both estimators are trivially parallelizable: the singular value decomposition computations for each subsample can be performed simultaneously and then added at the end. If B machines were available for the computations, the overall parallel computation time would be $O(mp^2)$ provided that $\log(B) \lesssim m$.

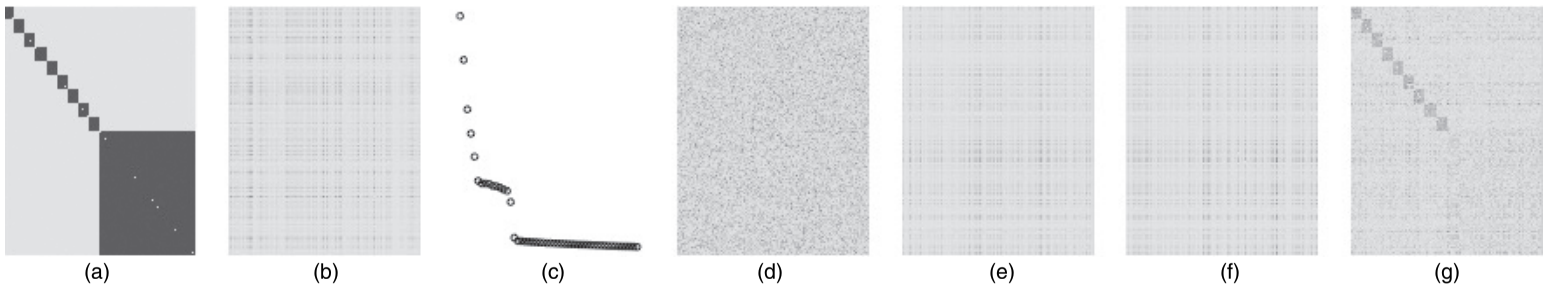

Fig. 1. Example with $p = 1000$ variables, sample size $n = 500$ and $q = 20$ latent confounders with strength $\nu = 0.5$, as described in detail in Section 5: (a) absolute values of the idiosyncratic covariance matrix Σ, showing block structure with different block sizes; (b) empirical covariance matrix $\hat{\Theta}$; (c) eigenvalues of $\hat{\Theta}$ on a log-scale; absolute values of PC removal estimator $\hat{\Sigma}(l)$, where l is chosen as (d) the oracle value $l = q$, the empirical estimators of q, (e) Hallin and Liška (2007) and (f) Bai and Ng (2002) suggested in the principal orthogonal complement thresholding method (Fan *et al.*, 2013); (g) proposed RSVP estimator with a subsample size of $m = 20$; RSVP manages to recover the smaller blocks, whereas the PC removal methods seemingly fail to recover any structure in the covariance matrix

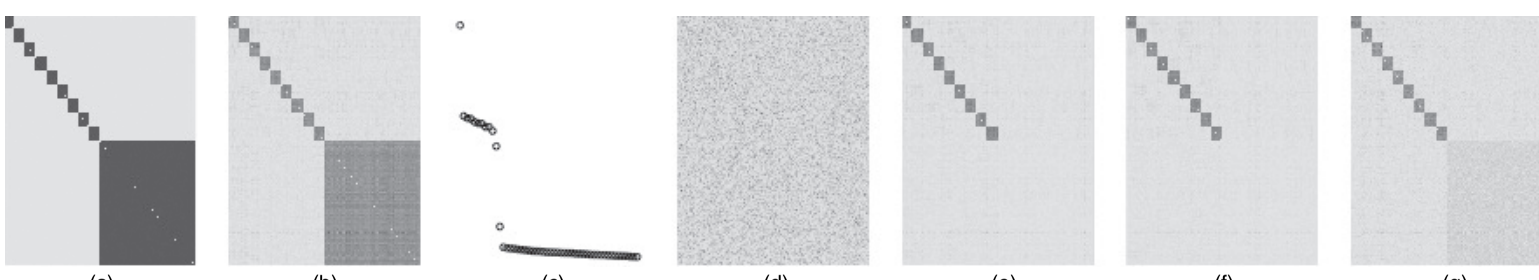

Fig. 2. Same setting as in Fig. 1 but for sample size $n = 1000$ and very weak latent confounding ($\nu = 0.01$): the large block is now even visible in the empirical covariance matrix; the PC-removal-based methods fail to recover the structure of the large block as they all remove at least one PC

2.5. Example

Figs 1 and 2 show an example of the proposed sample splitting RSVP estimator, compared with the target Σ and PC removal. In Fig. 1, the latent confounding is so strong that the empirical covariance shows very little visual indication of the block structure of the idiosyncratic covariance. Likewise, PC removal fails to recover the structure, whether we use an oracle for determining the number of factors to remove or estimate the optimal number of factors. RSVP in contrast recovers the smaller blocks. It is shown here for $m = 20$ samples in each subsample (default) but the results do not change appreciably when choosing a different subsample size. When reducing the strength of the latent confounders (Fig. 2), the empirical covariance shows the correct underlying structure visually but all PC removal methods fail to recover the largest block of variables as even just removing the first PC removes the large block.

3. Theoretical properties

In this section, we present some theoretical properties of $\hat{\Sigma}_{\text{rsvp}}$ and $\hat{\Sigma}_{\text{rsvp-split}}$. We first explain how $\hat{\Sigma}_{\text{rsvp}}$ has low variance and then argue that its bias is also well controlled in the high dimensional setting. We then discuss the consequences for $\hat{\Sigma}_{\text{rsvp-split}}$. We shall assume condition 1 below in several of the results to follow.

Condition 1. There exist constants $0 < c_1 < c_2$ such that $c_1 < \sigma_l \leqslant \max_j \text{var}(x_j) < c_2$. There exists a constant $c_3 > 1$ such that $\gamma_l > c_3 \sigma_u$ and $p > c_3 n$. Furthermore $\log(p) = o(n)$.

Theorem 1. Assume condition 1 and that $\sigma_u \lesssim p/\{n \log(p)\}$ and $q \lesssim n/\log(p)$. Then there exists a constant $c > 0$ such that, with probability at least $1 - c/p$,

$$\frac{p}{n} \|\hat{\Sigma}_{\text{rsvp}} - \mathbb{E}(\hat{\Sigma}_{\text{rsvp}})\|_\infty \lesssim \sqrt{\left\{\frac{\log(p)}{n}\right\}}.$$

We show in theorem 2 below that the entries in $\mathbb{E}(\hat{\Sigma}_{\text{rsvp}})$ are of the order n/p, so the result shows that the rate at which $\hat{\Sigma}_{\text{rsvp}}$ concentrates is equivalent to that enjoyed by the empirical covariance matrix in the absence of confounding. The proof, given in section E.2 in the online supplementary material, is based on a variant of the classical concentration inequality for a Lipschitz function $f : \mathbb{R}^d \to \mathbb{R}$ of IID Gaussian random variables $\zeta \sim \mathcal{N}_d(0, I)$, which may be of independent interest. Whereas the original result guarantees fast concentration when $\sup_{v \in \mathbb{R}^d} \|\nabla f(v)\|_2$ is small, our new result (theorem 7 in the supplementary material) requires only a high probability bound on $\|\nabla f(\zeta)\|_2$, and a potentially loose bound on $\mathbb{E}\|\nabla f(\zeta)\|_2^2$. See also lemma 1.3 of Klochkov and Zhivotovskiy (2018) for a related result.

Although our proof technique for concentration of $\hat{\Sigma}_{\text{rsvp}}$ makes use of particular properties of Gaussian distributions, one attractive feature of the estimator is that it enjoys a certain in-built robustness to deviations from Gaussianity in the distribution of X. Indeed, consider now the weaker requirement that

$$X = MZ\Theta^{1/2} + \mathbf{1}\mu^{\mathrm{T}} \tag{5}$$

where $M \in \mathbb{R}^{(n+1)\times(n+1)}$ is invertible and the rows of $Z \in \mathbb{R}^{(n+1)\times p}$ are independent following (potentially different) spherically symmetric distributions, so $ZQ =^{\mathrm{d}} Z$ for any orthogonal matrix $Q \in \mathbb{R}^{(n+1)\times(n+1)}$. A sufficient condition for this to occur is that the rows of X are IID and have a density with elliptical contours. In this more general setting we have the following result.

Proposition 3. The law of $\hat{\Sigma}_{\text{rsvp}}$ under requirement (5) above is the same as that when X has independent rows distributed as $\mathcal{N}_p(\mu, \Theta)$.

For example, the entries in Z can have arbitrarily heavy tails; provided that the spherical symmetry is satisfied, all results in this section hold under this setting and more generally under requirement (5). This may seem surprising at first sight but is analogous to how, if ζ has a spherically symmetric distribution, then the distribution of $\zeta/\|\zeta\|_2$ is simply the uniform distribution on the d-dimensional spherical shell, and in particular identical to the distribution that is obtained when $\zeta \sim \mathcal{N}_d(0, I)$.

We now turn to the expectation of $\hat{\Sigma}_{\text{rsvp}}$. Theorem 2 below shows that $\mathbb{E}(\hat{\Sigma}_{\text{rsvp}})$ is approximately a scaled version of Σ.

Theorem 2. Assume condition 1. We have that $\mathbb{E}(\hat{\Sigma}_{\text{rsvp}}) = PC^2P^{\mathrm{T}}$ where C is a diagonal matrix with C satisfying

$$\max_{j,k\in\{q+1,\ldots,p\}} \left| \frac{C_{jj}^2}{D_{jj}^2} - \frac{C_{kk}^2}{D_{kk}^2} \right| \lesssim \sigma_u \frac{n^2}{p^2}. \tag{6}$$

This result shows that the ratio of C_{jj}^2 to D_{jj}^2 does not vary much across $j \in \{q+1, \ldots, p\}$ provided that $p \gg n$. In fact we also have

$$\max_{j\in\{q+1,\ldots,p\}} \left| C_{jj}^2 - \frac{(n-q)D_{jj}^2}{\sum_{k=q+1}^{p} D_{kk}^2} \right| \lesssim \sqrt{\left(\frac{n}{p}\right)} + \frac{p}{\gamma_l n} \tag{7}$$

in the case where σ_u is bounded, which reveals the form of the scale factor, and in particular its dependence on the unknown q. A derivation is given in section F of the on-line supplementary material. We do not make direct use of this in the proof of theorem 3 below, however, as it is useful only when γ_l is large; in contrast, expression (6) is valid for any value of γ_l.

Combining the results of proposition 1 and theorems 1 and 2 gives the following high probability bound on the l_∞-norm error of estimating Σ, up to an unknown scale factor.

Theorem 3. Assume condition 1 and that $\sigma_u \lesssim p/\{n\log(p)\}$, $q \lesssim n/\log(p)$. With probability at least $1 - c/p$ for some constant $c > 0$, we have that there exists $\kappa > 0$ such that

$$\|\Sigma - \kappa\hat{\Sigma}_{\text{rsvp}}\|_\infty \lesssim \frac{\gamma_u\rho_1^2}{\gamma_l^2} + \rho_1\rho_2 + \min\left(\frac{p}{n}, \gamma_u\right)\rho_2^2 + \sigma_u\frac{n}{p} + \sqrt{\left\{\frac{\log(p)}{n}\right\}}. \tag{8}$$

If we additionally assume that $\rho_2^2 \lesssim q/p$ and ρ_1 is bounded, we have that there exists $\kappa > 0$ such that

$$\|\Sigma - \kappa\hat{\Sigma}_{\text{rsvp}}\|_\infty \lesssim \frac{\gamma_u}{\gamma_l^2} + \sqrt{\left(\frac{q}{p}\right)} + \frac{q}{n} + \sigma_u\frac{n}{p} + \sqrt{\left\{\frac{\log(p)}{n}\right\}}. \tag{9}$$

The first two terms in the bounds (8) and (9) come directly from the population level result proposition 1. The remaining terms do not depend on γ_l, demonstrating how RSVP, in contrast with the PC removal approach, does not rely on a large eigengap between $\Gamma^{\mathrm{T}}\Gamma$ and Σ. The final $\sqrt{\{\log(p)/n\}}$-term is due to the variance (see theorem 1). Considering expression (9), in the case where $\sigma_u \lesssim p\sqrt{\log(p)}/n^{3/2}$, $q \lesssim \sqrt{\{n\log(p)\}}$ and $p\sqrt{\log(p)} \geqslant n^{3/2}$, we have that with high probability

$$\inf_{\kappa>0} \|\Sigma - \kappa\hat{\Sigma}_{\text{rsvp}}\|_\infty \lesssim \frac{\gamma_u}{\gamma_l^2} + \sqrt{\left\{\frac{\log(p)}{n}\right\}}.$$

If the condition number of $\Gamma^{\mathrm{T}}\Gamma$ were bounded, we need only $\gamma_l \gtrsim \sqrt{\{n/\log(p)\}}$ for the l_∞-norm

error above to be of the same order as that achieved by the empirical covariance matrix of the (unobserved) unconfounded data W.

Although RSVP does not require strong eigengap conditions, we do need $p \gg n$ so that the term involving $\sigma_u n/p$, due to the bias of the estimator, is small. By sample splitting and averaging in constructing $\hat{\Sigma}_{\text{rsvp-split}}$, we effectively reduce n, but only introduce an extra $\sqrt{\log(p)}$-factor in the variance term, as the following result shows.

Theorem 4. Let $\hat{\Sigma}_{\text{rsvp-split}}$ be the sample splitting RSVP estimator with B subsamples of size m, so $n+1 = mB$. We consider for simplicity the case where the data are column centred in each subsample. Assume condition 1, but without the requirement that $c_3 n < p$; instead suppose that $c_1 m < p$ for $c_1 > 1$, and $B < p^{c_2}$ for some $c_2 > 0$. Assume that $1 \lesssim \sigma_u \lesssim p/\{m \log(p)\}$ and $q \lesssim m/\log(p)$. With probability at least $1 - c/p$ for some constant $c > 0$, we have that there exists $\kappa > 0$ such that

$$\|\Sigma - \kappa\hat{\Sigma}_{\text{rsvp-split}}\|_\infty \lesssim \frac{\gamma_u \rho_1^2}{\gamma_l^2} + \rho_1\rho_2 + \min\left(\frac{p}{m}, \gamma_u\right)\rho_2^2 + \sigma_u \frac{m}{p} + \frac{\log(p)}{\sqrt{n}}.$$

If we additionally assume that $\rho_2^2 \lesssim q/p$ and ρ_1 is bounded, we have that there exists $\kappa > 0$ such that

$$\|\Sigma - \kappa\hat{\Sigma}_{\text{rsvp-split}}\|_\infty \lesssim \frac{\gamma_u}{\gamma_l^2} + \sqrt{\left(\frac{q}{p}\right)} + \frac{q}{m} + \sigma_u \frac{m}{p} + \frac{\log(p)}{\sqrt{n}}. \tag{10}$$

Considering expression (10), we see that for an optimal $m \asymp \sqrt{(pq/\sigma_u)}$ we have with high probability that

$$\inf_{\kappa > 0} \|\Sigma - \kappa\hat{\Sigma}_{\text{rsvp-split}}\|_\infty \lesssim \frac{\gamma_u}{\gamma_l^2} + \sqrt{\left(\frac{q\sigma_u}{p}\right)} + \frac{\log(p)}{\sqrt{n}}. \tag{11}$$

Although the simple RSVP estimator is most useful in the high dimensional case $p \gg n$, this result shows that sample splitting gives good performance in moderate to low dimensional settings, which will be confirmed empirically in Section 5.

3.1. Weak confounding

The results and discussion thus far have considered the case where $\gamma_l > \sigma_u$. In cases where the confounding is sufficiently weak such that

$$\|\Theta - \Sigma\|_\infty = \|\Gamma\Gamma^{\mathrm{T}}\|_\infty = \max_j \|\Gamma^{\mathrm{T}}\Pi_\Gamma e_j\|_2^2 \leqslant \gamma_u \rho_2^2$$

is small, and so the empirical covariance $\hat{\Theta}$ is itself a good estimator of Σ, a straightforward consequence of our previous results and their proofs is that RSVP will behave similarly to the empirical covariance.

Corollary 1. Consider the set-up of theorem 3 but now without any restriction on q and γ_l (so in particular $\gamma_l < \sigma_u$ is permitted). With probability at least $1 - c/p$ for some constant $c > 0$, there exists $\kappa > 0$ such that

$$\|\Sigma - \kappa\hat{\Sigma}_{\text{rsvp}}\|_\infty \lesssim \gamma_u \rho_2^2 + (\gamma_u + \sigma_u)\frac{n}{p} + \sqrt{\left\{\frac{\log(p)}{n}\right\}}.$$

Suppose additionally that γ_u/γ_l is bounded, $\rho_2^2 \lesssim q/p$, $\sigma_u \lesssim p\sqrt{\log(p)}/n^{3/2}$,

$$q \lesssim n \max\{\sqrt{\{\log(p)/n\}}, 1/\log(p)\}$$

and $p\log(p) \geqslant n^2$. Then, with probability at least $1 - c/p$,

$$\|\Sigma - \kappa\hat{\Sigma}_{\text{rsvp}}\|_\infty \lesssim \sqrt{\left\{\frac{\log(p)}{n}\right\}}.$$

Note that the final result holds regardless of the strength of confounding, which can be arbitrarily weak or strong, though it relies on the condition number of $\Gamma^T\Gamma$ being bounded.

4. Conditional independence graph estimation and causal structure learning

In this section we consider using RSVP in conjunction with existing methods for conditional independence graph (CIG) estimation and causal structure learning. We first turn to the problem of estimating the CIG corresponding to Σ: this is the undirected graph on p nodes with an edge between nodes j and k with $j \neq k$ if and only if $w_j \not\perp\!\!\!\perp w_k | w_{-jk}$, where recall that $w \sim \mathcal{N}_p(\mu_w, \Sigma)$. Equivalently, we have an edge between j and k if and only if the precision matrix $\Omega = \Sigma^{-1}$ has $\Omega_{jk} \neq 0$.

4.1. Conditional independence graph estimation

Methods for CIG estimation when $p \gg n$ typically rely on Ω being sparse. Applying them directly to the observed data X will in general not work well, firstly as the inverse covariance Θ^{-1} of the observed data may be far from Ω, and secondly because Θ^{-1} will not be sparse but rather a sum of the sparse Ω and a low rank component due to the presence of latent confounding. However, many of the methods for sparse precision matrix estimation require only an estimated covariance as input and so can be readily applied to any estimate of Σ. Examples include neighbourhood selection (Meinshausen and Bühlmann, 2006), the graphical lasso (Yuan and Lin, 2007; Yuan, 2010; Friedman *et al.*, 2008) and constrained l_1-minimization for inverse matrix estimation (Cai *et al.*, 2011). Note that, as RSVP estimates Σ up to an unknown scale factor only, we can similarly only hope to recover the precision matrix up to an unknown scale factor; this, however, suffices for estimating the CIG. Theoretical results for constrained l_1-minimization for inverse matrix estimation and the graphical lasso require only an initial estimate of Σ that is close in l_∞-norm, so our estimation error bounds for Σ translate directly into estimation error bounds on Σ^{-1}. We now present the corresponding result for neighbourhood selection, which is more delicate.

The procedure of neighbourhood selection involves running p lasso regressions of each variable against all others. The resulting coefficient estimates may then be used to derive an estimate of the CIG. Phrased in terms of an estimate $\hat{\Sigma}$ of the covariance, the so-called nodewise regressions take the form

$$\hat{\beta}^{(j)} := \underset{b \in \mathbb{R}^p : b_j = 0}{\arg\min} \left\{ \tfrac{1}{2} b^T \hat{\Sigma} b - b^T \hat{\Sigma}_j + \lambda_j \|b\|_1 \right\}. \tag{12}$$

The population level minimizer $\beta^{(j)} \in \mathbb{R}^p$ (i.e. with $\hat{\Sigma}$ replaced by Σ) of the above when $\lambda_j = 0$ satisfies

$$\beta_l^{(j)} = \begin{cases} (\Sigma_{-j,-j}^{-1} \Sigma_{-j,j})_l & l < j, \\ 0 & l = j, \\ (\Sigma_{-j,-j}^{-1} \Sigma_{-j,j})_{l-1} & l > j. \end{cases}$$

The $\{\beta^{(j)}\}_{j=1}^p$ encode the CIG; indeed $w_j \perp\!\!\!\perp w_k | w_{-jk}$ if and only if $\beta_k^{(j)} = \beta_j^{(k)} = 0$. Here we shall take $\hat{\Sigma} = \hat{\Sigma}_{\text{rsvp}}$ in expression (12); we thus expect that a scaled version of $\hat{\beta}^{(j)}$ becomes close to $\beta^{(j)}$.

To present our result on the statistical properties of $\hat{\beta}^{(j)}$, we introduce the following quantities. Let $S_j = \{l : \beta_l^{(j)} \neq 0\}$ and let $s_j := |S_j|$ and $s = \max_j s_j$; thus s_j and s are the degree of the jth node and the maximal degree in the CIG respectively. Also define

$$\eta_j = (\beta^{(j)})^{\mathrm{T}} \Gamma \Gamma^{\mathrm{T}} \beta^{(j)}.$$

Our theory will require the η_j to be small. We always have $\eta_j \lesssim s_j$ for all j. Indeed

$$(\beta^{(j)})^{\mathrm{T}} \Gamma \Gamma^{\mathrm{T}} \beta^{(j)} \leqslant (\beta^{(j)})^{\mathrm{T}} \Theta \beta^{(j)} = \beta^{(j)}_{S_j} \Theta_{S_j, S_j} \beta^{(j)}_{S_j}.$$

As Θ is positive semidefinite, we have $|\Theta_{lk}| \leqslant \max(\Theta_{ll}, \Theta_{kk}) \lesssim 1$ for all l and k. Thus, by the Gershgorin circle theorem, $\lambda_{\max}(\Theta_{S_j, S_j}) \lesssim s_j$. Also, as

$$1 \gtrsim \Theta_{jj} \geqslant \mathrm{var}(w_j) \geqslant \mathrm{var}(w_j | w_{-j}) = \| \Sigma^{-1/2}_{-j,-j} \Sigma_{-j,j} \|_2^2 \geqslant \sigma_l \| \Sigma^{-1}_{-j,-j} \Sigma_{-j,j} \|_2^2,$$

we have $\| \beta^{(j)} \|_2 \lesssim 1$, whence $\eta_j \lesssim s_j$.

However, in many settings we can expect the η_j to be smaller: if we consider the column space of Γ to have been chosen (by nature) uniformly at random conditionally on Σ, then we have

$$\eta_j \lesssim 1 \qquad \text{for all } j. \tag{13}$$

A derivation of this is given in section I of the on-line supplementary material.

Theorem 5. Assume condition 1. Let

$$\Delta := \sqrt{\left\{ \frac{sn}{\log(p)} \right\}} \left\{ \frac{\gamma_u \rho_1^2}{\gamma_l^2} + \rho_1 \rho_2 + \min\left(\frac{p}{n}, \gamma_u \right) \rho_2^2 + \sigma_u \frac{n}{p} \right\}.$$

Let $\hat{\beta}^{(j)}$ be the nodewise regression coefficient when $\hat{\Sigma} = \hat{\Sigma}_{\mathrm{rsvp}}$ and

$$\lambda_j = A \sqrt{\{ \max(\eta_j, \Delta, 1) n \log(p) \} / p}$$

for constant $A > 0$. Suppose that $s = o[\sqrt{\{n / \log(p)\}}]$. We have that, for A, n and p sufficiently large, with probability at least $1 - c/p$ for some constant $c > 0$,

$$\| \hat{\beta}^{(j)} - \beta^{(j)} \|_2 \lesssim \sqrt{\{ s_j \log(p) \max(\eta_j, \Delta, 1) / n \}},$$
$$\| \hat{\beta}^{(j)} - \beta^{(j)} \|_1 \lesssim s_j \sqrt{\{ \log(p) \max(\eta_j, \Delta, 1) / n \}}$$

for all $j = 1, \ldots, p$.

Suppose that $\rho_2^2 \lesssim q/p$, $\rho_1 \lesssim 1$, $\gamma_l^2 / \gamma_u \gtrsim \sqrt{\{sn / \log(p)\}}$, $q \lesssim \sqrt{\{n \log(p)/s\}}$ and $\sigma_u \lesssim p \sqrt{\{\log(p)/(sn^3)\}}$; then $\Delta \lesssim 1$. If in addition $\eta_j \lesssim 1$ for all j, we recover the usual estimation error rates for the lasso:

$$\| \hat{\beta}^{(j)} - \beta^{(j)} \|_2 \lesssim \sqrt{\{ s_j \log(p)/n \}},$$
$$\| \hat{\beta}^{(j)} - \beta^{(j)} \|_1 \lesssim s_j \sqrt{\{ \log(p)/n \}}.$$

The following simple corollary shows that, under a minimum signal strength condition, appropriately thresholding the estimates $\hat{\beta}^{(j)}$ recovers the true CIG.

Corollary 2. Consider the set-up of theorem 5 and suppose $\max(\Delta, \eta_j)$. Suppose that

$$\min_{k \in S_j} |\beta^{(j)}_k| \geqslant C \sqrt{\{ s_j \log(p)/n \}}$$

for all j and some $C > 0$. For C sufficiently large, with probability at least $1 - c/p$ for some $c > 0$, there exists $\tau > 0$ such that, defining

$$\hat{S}_j = \{ k : |\hat{\beta}^{(j)}_k| \geqslant \tau \sqrt{\{ s_j \log(p)/n \}} \},$$

we have $\hat{S}_j = S_j$ for all j.

Although edges in a CIG are typically given a causal interpretation, structural equation models (Pearl, 2009) and graphical modelling with directed acyclic graphs (Lauritzen, 1996) offer a more principled approach for causal inference. Below we explain how the popular PC algorithm (Spirtes *et al.*, 2000) may be run with our RSVP estimate as its input to enable causal structure learning in the presence of hidden confounding.

4.2. Causal structure learning

In this section we describe how our RSVP estimator may be used for causal structure learning concerning the unconfounded $w \sim \mathcal{N}_p(\mu_w, \Sigma)$. If we assume a structural causal model for w with an underlying *directed acyclic graph* (DAG) encoding parent–children relationships (Pearl, 2009), then the observational distribution factorizes according to this DAG. The interventional distributions under 'do' interventions can then be obtained by truncated factorizations (Robins, 1986; Pearl, 2009) under an assumption known as autonomy (Haavelmo, 1944).

If the underlying DAG G is unknown, it needs to be estimated from data; for a general overview of causal structure learning see for example Heinze-Deml *et al.* (2018). Under a faithfulness assumption (Meek, 1995), the set of conditional independences in the observational distribution will be exactly those that may be inferred via d-separation from G. In general, many DAGs will be compatible with the observational distribution in this way and these form an equivalence class which may be conveniently represented through a *completed partially directed acyclic graph* (CPDAG). A CPDAG contains both directed and undirected edges, and essentially contains all the information relating to causal structure that may be inferred from a given observational distribution under the assumption of faithfulness.

Our goal here is to infer the CPDAG corresponding to the distribution of the unconfounded data. To do this we employ the PC algorithm (Spirtes *et al.*, 2000; Kalisch and Bühlmann, 2007). The population version of the PC algorithm is a procedure for determining the CPDAG $C(G)$ corresponding to a distribution P that is faithful to a DAG G given a list of conditional independences satisfied by P. In our context where $P = \mathcal{N}_p(\mu_w, \Sigma)$ with Σ positive definite, these conditional independences may be equivalently represented by partial correlations: we have for $w \sim \mathcal{N}_p(\mu_w, \Sigma)$ that

$$w_j \perp\!\!\!\perp w_k | w_S \Leftrightarrow \rho_{jk|S} = 0, \tag{14}$$

where the partial correlation $\rho_{jk|S}$ satisfies

$$\rho_{jk|S} = -\frac{\Psi_{jk}}{\sqrt{(\Psi_{uu}\Psi_{vv})}}, \tag{15}$$

and $\Psi^{-1} = \Sigma_{A,A}$ with $A = \{j\} \cup \{k\} \cup S$ (Harris and Drton, 2013). Here we have indexed the rows and columns of Ψ according to the elements of A.

The sample version of the PC algorithm replaces queries of conditional independence with conditional independence tests. In our case, in analogy with expressions (14) and (15) we shall consider tests that declare the conditional dependence $w_j \not\perp\!\!\!\perp w_k | w_S$ if and only if

$$\frac{|\hat{\Psi}_{jk}|}{\sqrt{(\hat{\Psi}_{uu}\hat{\Psi}_{vv})}} \geqslant \tau, \tag{16}$$

where $\hat{\Psi}^{-1} = \hat{\Sigma}_{A,A}$ with $\hat{\Sigma}$ either $\hat{\Sigma}_{\text{rsvp}}$ or $\hat{\Sigma}_{\text{rsvp-split}}$; here A is defined as above and threshold τ is a tuning parameter. If $\hat{\Sigma}_{A,A}$ is not invertible, we shall simply accept the null of conditional independence.

In the case where confounding is not present, the PC algorithm requires faithfulness and a certain minimum signal strength condition for partial correlations. We shall therefore assume that $\mathcal{N}_p(\mu_w, \Sigma)$ is faithful to a DAG G and our target of inference will be the corresponding CPDAG $C(G) =: C$. We denote the maximum degree of G by d. Define also the following parameter controlling minimum signal strength:

$$\omega := \min\{|\rho_{jk|S}| : j, k \in V, S \subseteq V, |S| \leqslant d, \rho_{jk|S} \neq 0\}.$$

It will also be convenient to introduce a particular minimum restricted eigenvalue σ_r of Σ defined through $\sigma_r := \min_{I:|I|\leqslant d+2} \lambda_{\min}(\Sigma_{I,I})$. Note that we always have $\sigma_r \geqslant \sigma_l$.

The result below follows directly from the proof of theorem 8 in Harris and Drton (2013).

Lemma 1. Let $\hat{C}_\tau$ be the output of the PC algorithm using conditional independence tests given by expression (16) with threshold τ. For any $A \geqslant 1$ we have

$$\mathbb{P}\left(\hat{C}_\tau = C \text{ for all } \tau \in \left[\frac{\omega}{2A}, 1 - \frac{\omega}{2A}\right]\right) \geqslant \mathbb{P}\left\{\inf_{\kappa>0} \|\kappa\hat{\Sigma} - \Sigma\|_\infty \leqslant \frac{\omega\sigma_r^2}{(4A + \omega + \sigma_r\omega)(d+2)}\right\}.$$

Taking $\hat{\Sigma}$ as either $\hat{\Sigma}_{\text{rsvp}}$ or its sample splitting variant $\hat{\Sigma}_{\text{rsvp-split}}$, by combining lemma 1 with one of theorems 3 or 4, we can obtain high probability guarantees on recovering the CPDAG corresponding to the unconfounded data. As an example, we consider the setting where the assumptions of theorem 4 and those leading to expression (11) hold. Additionally, consider an asymptotic regime where

$$\frac{\sigma_u}{\sigma_l^2} + \sqrt{\left(\frac{q\sigma_u}{p}\right)} + \frac{\log(p)}{\sqrt{n}} = o\left(\frac{\omega}{d}\right). \tag{17}$$

Then using $\hat{\Sigma}_{\text{rsvp-split}}$ with an optimal subsample size $m \asymp \sqrt{(pq/\sigma_u)}$ we have the following conclusion: there is a sequence $a_n \to 0$ and constant $c > 0$ such that $\hat{C}_\tau = C$ for all $\tau \in [a_n, 1 - a_n]$ with probability at least $1 - c/p$. We may compare this conclusion with the results that were obtained in Kalisch and Bühlmann (2007) that provide similar guarantees for the PC algorithm when confounding is not present. If we assume that the final term of $\log(p)/\sqrt{n}$ on the left-hand side of equation (17) is the dominant term, our requirement is $\log(p)/\sqrt{n} = o(\omega/d)$ whereas the equivalent result in Kalisch and Bühlmann (2007) requires only $\sqrt{\{\log(p)/n\}} = o(\omega/\sqrt{d})$. In particular, we see that, in our setting, the maximal degree d cannot grow as quickly. This restriction is also present in the analogous result of Harris and Drton (2013) who considered applying the PC algorithm (in the absence of hidden confounding) by using conditional independence tests based on partial correlations derived from rank correlations.

5. Numerical results

5.1. Simulation experiments

In this section we provide some numerical results for various scenarios and compare the proposed estimator with the PC removal estimators, as employed in the principal orthogonal complement thresholding method (Fan *et al.*, 2013). Results for shrinkage estimators of Ledoit–Wolf type (Ledoit and Wolf, 2004) are also included in our comparison.

5.1.1. Experimental set-ups

We consider five scenarios described below. For each of these, we generate $n \in \{100, 200, 500, 1000, 2000\}$ independent samples from $\mathcal{N}_p(0, \Theta)$ for a covariance matrix $\Theta \in \mathbb{R}^{p\times p}$ that has an idiosyncratic component and a component due to confounding $\Theta = \Sigma + \Gamma^{\mathrm{T}}\Gamma$. The number

of variables is varied in $p \in \{100, 200, 500, 1000, 2000\}$. For q latent confounders, the entries of the matrix $\Gamma \in \mathbb{R}^{p \times q}$ are sampled independently from a standard normal distribution, and column $k \in \{1, \ldots, q\}$ of Γ is scaled by a factor $\nu \exp(-k)$ to have a decaying spectrum among the latent confounders. The strength $\nu \in \{0.01, 0.1, 0.5, 1, 5, 20\}$ allows for a variation in the overall strength of the latent confounding. The five scenarios that are considered distinguish themselves by a different structure of the idiosyncratic covariance matrix Σ and the number of latent confounders q. All diagonal entries of Σ are set to 1.

(a) *Block structure*: the p variables are divided into 10 blocks of equal size. The correlation within each block is set uniformly to 0.95 and 0 outside blocks, with unit variance for all variables. There are $q = 20$ latent variables in this scenario.
(b) *Block structure II*: half of the variables are divided into 10 blocks of equal size, similarly to the previous scenario. The remaining variables form one large block. The within-block correlation is 0.5 and between-block correlation is again 0. The correlation within each block is set to 0.95 and all variables have unit variance. There are $q = 20$ latent variables in this scenario.
(c) *Toeplitz structure*: the inverse idiosyncratic covariance marix is set to a unit diagonal and first off-diagonal entries equal to -0.4999 (with circular extension). Variables are then scaled to have unit variance. There are again $q = 20$ latent variables in this scenario.
(d) *Toeplitz structure II*: the second Toeplitz design is identical to the previous Toeplitz design, except that the number of latent confounders is reduced to $q = 3$.
(e) *Erdős–Rényi structure*: the non-zero entries of the inverse idiosyncratic covariance are chosen randomly, each edge being selected with probability $10/p$. The diagonal of the inverse is set to unit values initially, and all off-diagonal entries are set to constants such that the sum of all non-diagonal entries in each row is bounded by 0.99 and the inverse matrix is hence diagonal dominant and invertible. The variables are in a second step again scaled to have unit diagonal entries in the idiosyncratic covariance Σ.

Varying the structure, number of samples n, dimension p and strength ν of the latent confounders, we run 200 simulations of each unique parameter configuration and compute the following quantities.

(a) We obtain the estimated covariance matrix $\hat{\Sigma}_{\mathrm{pca}}(l)$, where the number l is chosen first as $l = 0$, leading to the empirical covariance matrix. This first estimator is also the basis for comparisons with Ledoit–Wolf-type shrinkage (Ledoit and Wolf, 2004). (The results for a Ledoit–Wolf covariance estimator with the identity matrix as the shrinkage target are identical to those for PC removal with $l = 0$ (i.e. the empirical covariance matrix) as the objective that we measure will be unchanged by the shrinkage.) Next we use the oracle value $l = q$ (which is of course unavailable in practice) and then, as suggested in Fan *et al.* (2013), the values of the two estimators of q that are based on the respective first information criterion in Bai and Ng (2002) and Hallin and Liška (2007).
(b) We calculate the sample splitting RSVP estimator $\hat{\Sigma}_{\text{rsvp-split}}$ for subsample size $m \in \{20, 50, 70\}$.

Some results are shown in Figs 3 and 4. Other possible approaches such as the sparse–dense decomposition approach of Chandrasekaran *et al.* (2012) are unfortunately computationally infeasible for these settings.

We would like to compare for each estimate its accuracy with respect to the true idiosyncratic covariance in a suitable norm, which we chose here for simplicity as the Frobenius norm. To be invariant with respect to scaling, we may consider

Fig. 3. Example of block structure with $p = 1000$ and $n = 100$ and strong latent confounders ($\nu = 20$): the results are presented as in Fig. 1; the empirical covariance matrix $\hat{\Theta}$ and the PC removal estimates fail to recover the block structure

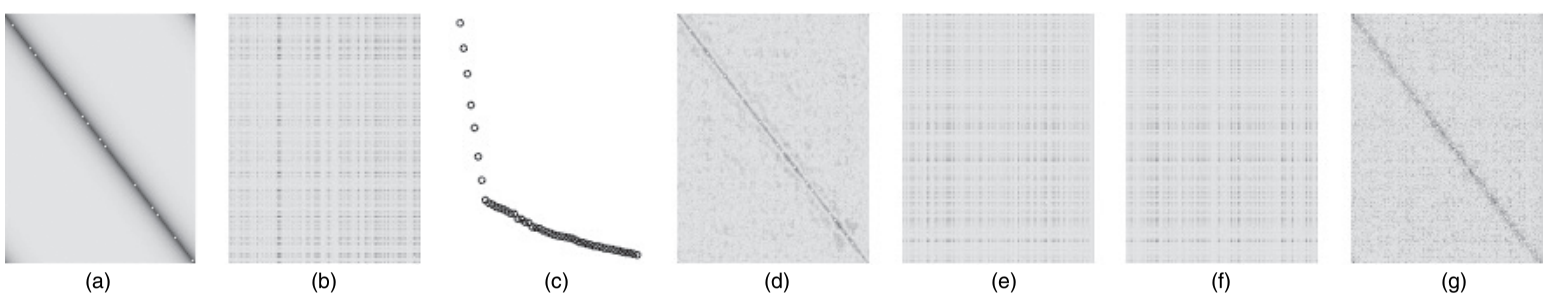

Fig. 4. Same set-up as in Fig. 3 for a Toeplitz stucture of the idiosyncratic covariance and $q = 20$ latent confounders

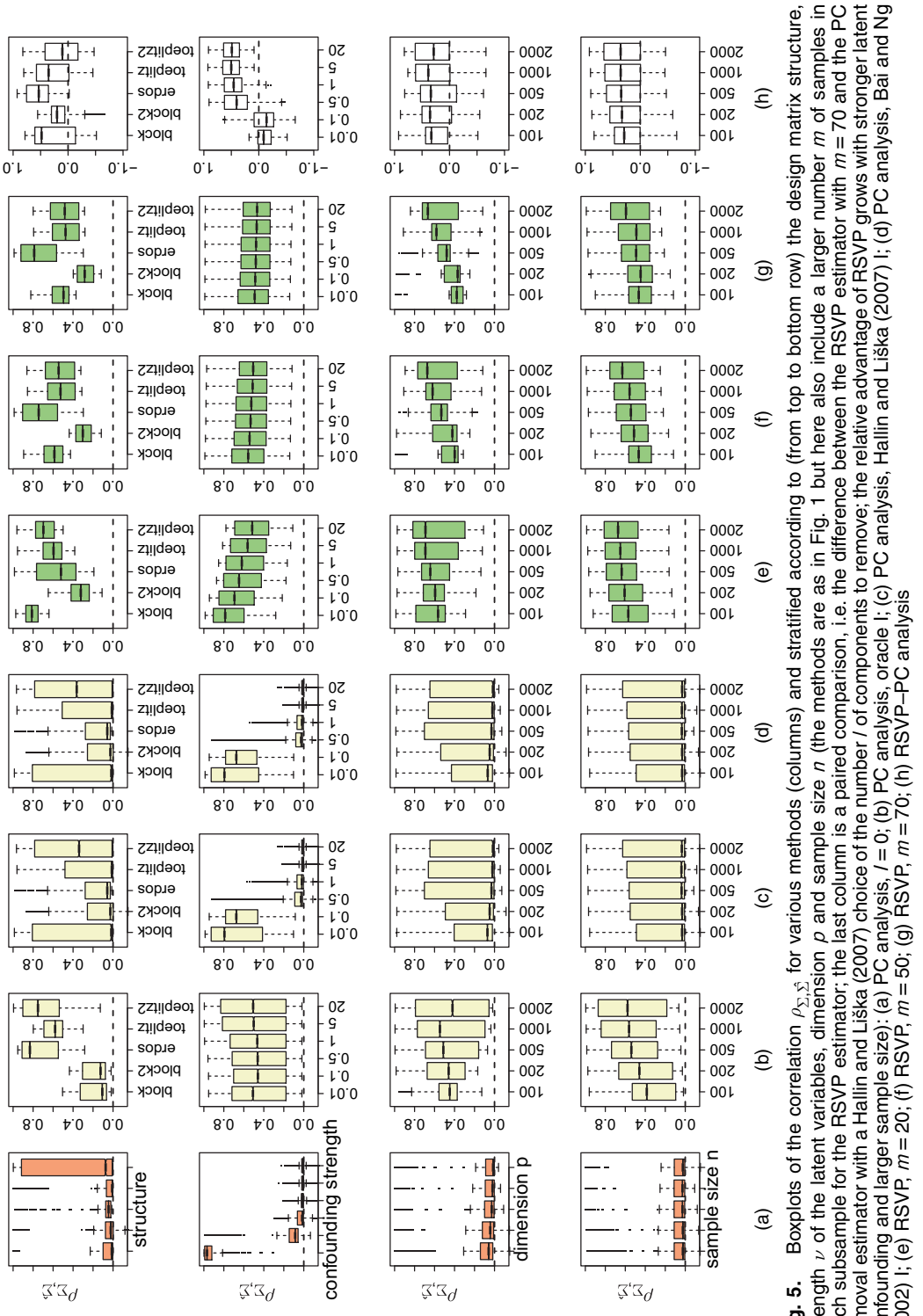

Fig. 5. Boxplots of the correlation $\rho_{\Sigma,\hat{\Sigma}}$ for various methods (columns) and stratified according to (from top to bottom row) the design matrix structure, strength ν of the latent variables, dimension p and sample size n (the methods are as in Fig. 1 but here also include a larger number m of samples in each subsample for the RSVP estimator; the last column is a paired comparison, i.e. the difference between the RSVP estimator with $m = 70$ and the PC removal estimator with a Hallin and Liška (2007) choice of the number l of components to remove; the relative advantage of RSVP grows with stronger latent confounding and larger sample size): (a) PC analysis, $l = 0$; (b) PC analysis, oracle l; (c) PC analysis, Hallin and Liška (2007) l; (d) PC analysis, Bai and Ng (2002) l; (e) RSVP, $m = 20$; (f) RSVP, $m = 50$; (g) RSVP, $m = 70$; (h) RSVP–PC analysis

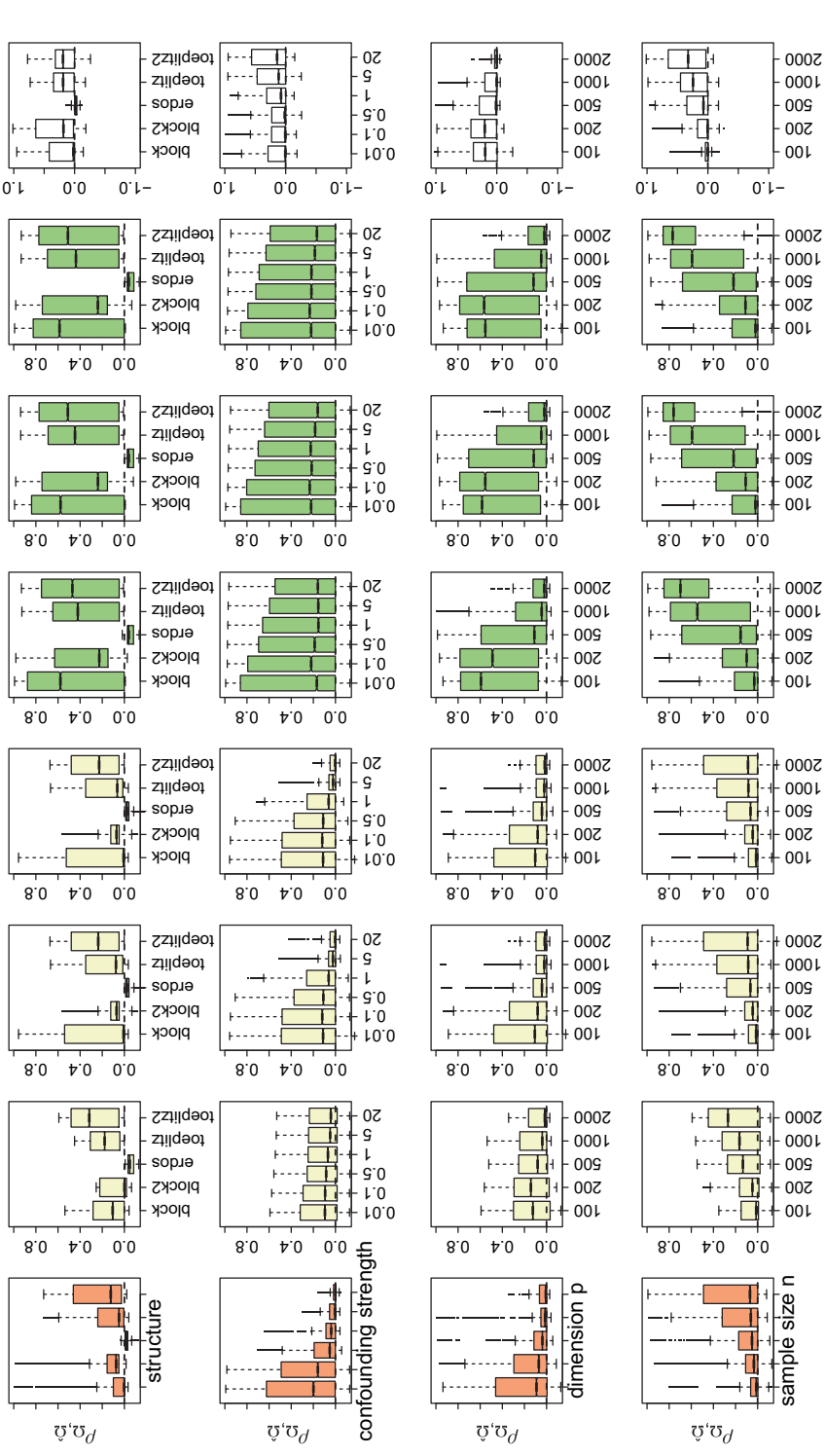

Fig. 6. Analogous results to those in Fig. 5 for inverse covariance matrix estimation (as before, the relative advantage of RSVP grows with stronger latent confounding and larger sample size): (a) PC analysis, $l = 0$; (b) PC analysis, oracle l; (c) PC analysis, Hallin and Liška (2007) l; (d) PC analysis, Bai and Ng (2002) l; (e) RSVP, $m = 20$; (f) RSVP, $m = 50$; (g) RSVP, $m = 70$; (h) RSVP–PC analysis

$$\inf_{\kappa>0} \|\Sigma - \kappa\hat{\Sigma}\|_{\mathrm{F}},$$

which is monotonically decreasing with the empirical correlation $\rho_{\Sigma,\hat{\Sigma}}$ between the vectorized matrices Σ and $\hat{\Sigma}$; we shall use $\rho_{\Sigma,\hat{\Sigma}}$ as a criterion for simplicity, and also we omit the diagonals from Σ and $\hat{\Sigma}$ in the computation. For estimation of the *inverse* covariance $\Omega := \Sigma^{-1}$, we invert our estimates above by using the approach of Meinshausen and Bühlmann (2006) as implemented in the R package `glasso` (Friedman *et al.*, 2018) to give estimates $\hat{\Omega}$. The penalty parameter is set to a very small uniform value of $\lambda = 10^{-6}$ for computational speed and easier comparison between methods. Cross-validation of the penalty is also not straightforward to implement here as we do not have access to clean data that would be free of the influence of the latent confounders.

5.1.2. *Results*

A summary of results from each of the $750 = 5 \times 5 \times 6 \times 5$ unique parameter settings is shown in Fig. 5. The RSVP estimator with low number $m = 20$ of samples in each subsample in general dominates the other estimators (in terms of having higher mean correlation and higher quartiles), no matter whether we stratify according to design matrix structure, strength of latent confounders, sample size or dimension of the graph. The only exception seems to be the case $\nu = 0.01$, where the latent confounders are effectively absent. Here the empirical covariance improves the RSVP estimator, as expected.

Comparing the various PC removal approaches, it is noteworthy that, for an increasing strength of the latent confounding, the oracle (true) value of q performs much better than using any of the suggested empirical estimates of q. In contrast, for weak confounding, removing all q latent confounders performs worse in general because of the decaying spectrum of the latent confounding: too much of the idiosyncratic covariance is removed by the oracle estimate in these cases. RSVP tends to perform at least as well as the optimal approach among the three PC removal approaches across all strengths of the latent confounding, even though in practice the oracle choice of q for PC removal is clearly not even available.

Analogous results for inverse covariance matrix estimation are shown in Fig. 6, with a single example outcome in Fig. 7. The differences between RSVP with different numbers of samples in each subsample are smaller, arguably because the error that is introduced by matrix inversion dominates the relatively small differences. Although estimating the covariance of a random Erdős–Rényi graph seems easy for the covariance, it becomes relatively difficult for the inverse covariance matrix. Finally, whereas a dimension of $p = 2000$ still yields very good results in Frobenius norm for covariance estimation, it seems to become very challenging for inverse covariance estimation.

The relative performance of the sample splitting version of RSVP as a function of number of samples m in each subsample is shown in Fig. 8. For very weak latent confounding, taking very small values of m performs optimally as the sampling splitting RSVP estimator then converges to the empirical covariance matrix. Whereas the scaling of the optimal m as proportional to $\sqrt{(pq)}/\sigma_u$ emerges from the theory, in our examples the choice $m = 2\sqrt{p}$ seems to be a good rule of thumb choice for the size of the subsamples.

5.1.3. *Model violations*

To investigate robustness against model violations for covariance estimation, we replace the normal distribution for the idiosyncratic noise for X and H by multivariate t-distributions with df_1 degrees of freedom, where $\mathrm{df}_1 \in \{1, 2, 3, 5, 10, 20, 50, 100\}$. We also generate the loading matrix Γ by using a multivariate t-distributions with df_2 degrees of freedom and we vary this

Fig. 7. Example of block structure with $p = 100$ and $n = 500$ and medium strong latent confounders ($\nu = 20$) (the results presented are analogous to those in Fig. 1, but here we are interested in inverse covariance estimation (via nodewise regression on the estimated idiosyncratic covariances)): (a) absolute values of Σ^{-1}; (b) empirical covariance matrix $\hat{\Theta}$; (c) eigenvalues of $\hat{\Theta}$ on a log-scale; (d)–(g) absolute values of the inverted estimated $\hat{\Sigma}$, using the same methods as in Fig. 1

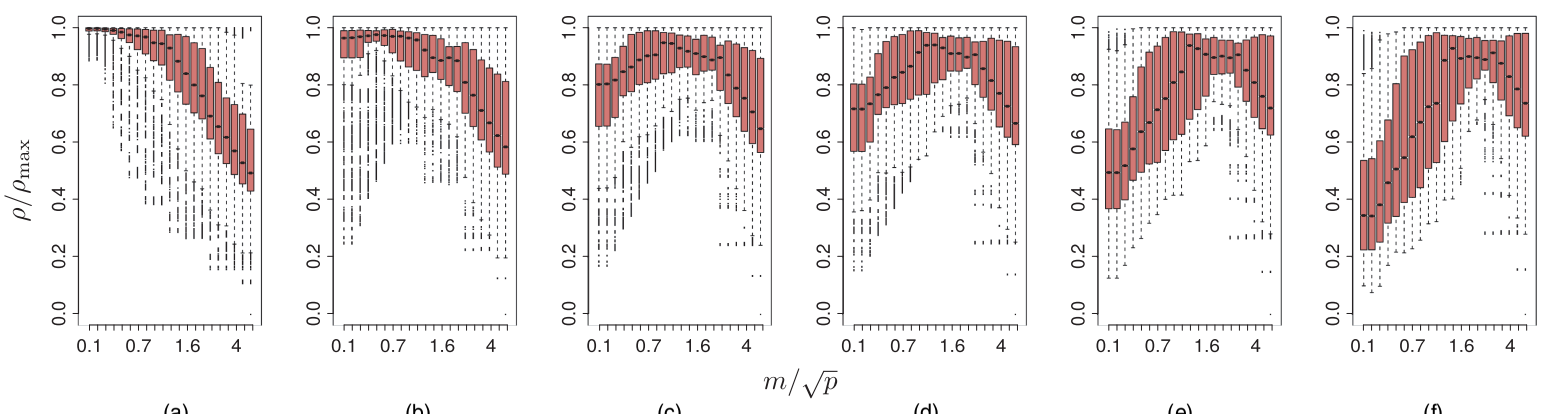

Fig. 8. Performance as a function of sample size m used for the sample splitting version of RSVP (for each scenario, we divide $\rho_{\Sigma,\hat{\Sigma}}$ by the maximal value across all subsample sizes m; the figure shows the boxplots for different strengths ν of latent confounding as a function of $m/\sqrt{p}$; for weak latent confounding (a) taking $m = 1$ is optimal as then RSVP is equivalent to the empirical covariance matrix on row-normalized data; for stronger latent confounding, a value $m = c\sqrt{p}$ with $c \approx 2$ performs well across a wide range of scenarios): (a) $\nu = 0$; (b) $\nu = 0.1$; (c) $\nu = 0.5$; (d) $\nu = 1$; (e) $\nu = 5$; (f) $\nu = 20$

parameter among the same set of values as those used for df_1. Analogously to Fig. 5, Fig. 9 shows the performance for covariance estimation marginally as a function of both df_1 and df_2, where the remaining parameters (graph structure, dimension, sample size and strength of confounding) are averaged out. The cases $\mathrm{df}_1 = 1$ and $\mathrm{df}_2 = 1$ correspond to Cauchy distributions. We comment here that Σ does not correspond to a covariance if $\mathrm{df}_1 \leqslant 2$. Nevertheless, Σ can still be identifiable from the distribution of w.

As an additional test of robustness, we consider, in a second set of experiments, replacing the linear structural equation $x = w + \Gamma h$ (1) with a max-linear model (Gissibl and Klüppelberg, 2018)

$$x_j = \max\{w_j, (\Gamma h)_j\}; \tag{18}$$

our goal is as before to recover $\Sigma = \mathrm{cov}(w)$. We present in Fig. 10 the results averaged over all other parameters of our simulation set-up (graph structure, dimension, sample size, strength of confounders, df_1 and df_2). The performances of both oracle PC removal and RSVP suffer in the max-linear case and drop to similar levels to the data-driven PC removal methods. However, even in this case, RSVP outperforms data-driven PC removal approaches; in the case of the Hallin and Liška (2007) choice of the number of components, RSVP gives better results in more than three-quarters of all simulation settings, as can be seen in Fig. 10(h).

5.2. GTEX data analysis

In this section we illustrate the key properties of RSVP on a collection of gene expression data sets made publicly available by the GTEX consortium (Aguet *et al.*, 2016). Such data sets are particularly prone to the type of confounding that is studied in this paper (Leek and Storey, 2007; Stegle *et al.*, 2012; Gagnon-Bartsch *et al.*, 2013). Our aim is to determine which genes are biologically related in that they regulate each other. To validate our results, we use the gene ontology database (Ashburner *et al.*, 2000).

The GTEX consortium conducted a large-scale ribonucleic acid sequencing experiment which resulted in the the collection of gene expression data from hundreds of donors in more than 50 human tissues. To carry out their analyses, they estimated confounders by leveraging external information such as gender and genetic relatedness between donors, and by inferring some confounders from the data themselves by using probabilistic estimation of expression residuals (PEER) (Stegle *et al.*, 2012). Both the confounders and the fully processed, normalized and filtered gene expression data are available on the website of the consortium (`https://gtexportal.org/home/datasets`; in addition, code to compute RSVP and subsampling versions, and also to reproduce all the results that are described in this section, is available from `https://github.com/benjaminfrot/RSVP`).

For each tissue $\mathcal{T}$, where $\mathcal{T}$ is for example whole blood, lung or thyroid, there is available a data matrix $X_{\mathcal{T}}$ of gene expression levels with dimensions $n_{\mathcal{T}} \times p_{\mathcal{T}}$ along with an $n_{\mathcal{T}} \times q_{\mathcal{T}}$ matrix of confounders. We removed tissues for which $n_{\mathcal{T}} \leqslant 100$; the 44 remaining tissues had a ratio $n_{\mathcal{T}}/p_{\mathcal{T}}$ ranging between 0.006 and 0.03 and values of $p_{\mathcal{T}}$ ranging between 14337 and 17855. (The list of tissues as well as the number of samples and variables for each of them can be found in the on-line supplementary materials.) In line with the analysis methods of the GTEX consortium, we used all the PEER factors at our disposal, resulting in a total number of $q_{\mathcal{T}}$ confounders for each tissue equal to the number of PEER factors for that tissue plus five confounders derived from external sources (e.g. donors' genotypes, gender, etc.). (According to the analysis methods section of the consortium's website 'the number of PEER factors was determined as a function of sample size N: 15 factors for $N < 150$, 30 factors for $150 \leqslant N < 250$,

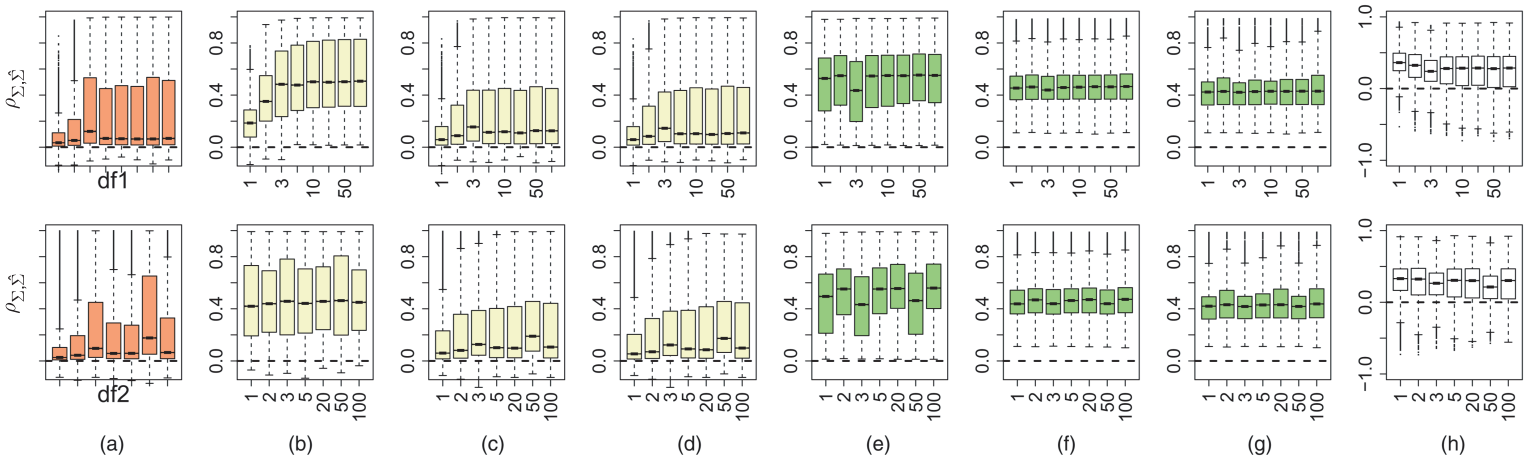

Fig. 9. Performance of PC removal methods and RSVP when replacing multivariate normal distributions for w and h, and generation mechanism for the entries of Γ by multivariate t-distributions with df_1 and df_2 degrees of freedom respectively (whereas the performance of the PC removal approaches deteriorates considerably for small degrees of freedom in the idiosyncratic noise distributions, the performance of RSVP is largely unaffected by these heavy-tailed distributions): (a) PC analysis, $l = 0$; (b) PC analysis, oracle l; (c) PC analysis, Hallin and Liška (2007); (d) PC analysis, Bai and Ng (2002) I; (e) RSVP, $m = 20$; (f) RSVP, $m = 50$; (g) RSVP, $m = 70$; (h) RSVP–PCA

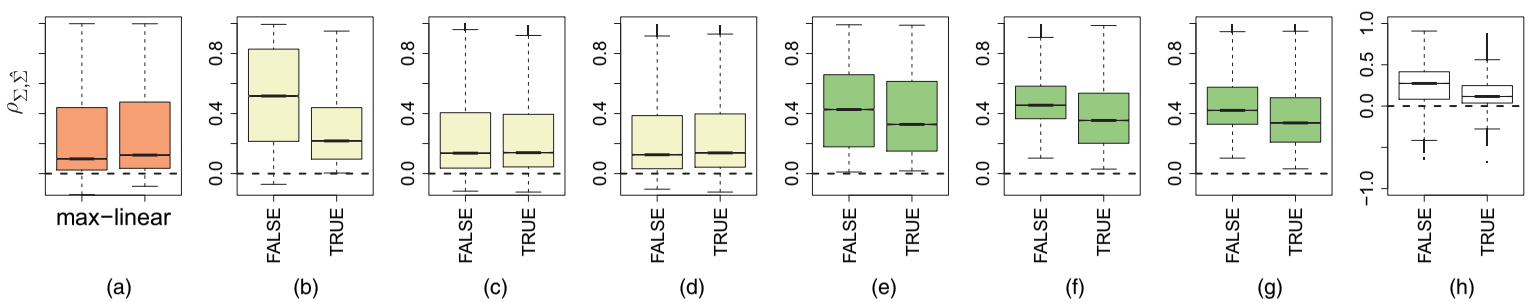

Fig. 10. Performance of PC removal methods and RSVP when the linear structural equation for x is replaced with a max-linear model (18) (the average performance for the linear model is shown as a boxplot for max-linear equal to 'false', whereas the max-linear case corresponds to the boxplot with max-linear equal to 'true'; we see that the advantage of RSVP over PC removal methods deteriorates under the max-linear model): (a) PC analysis, $l = 0$; (b) PC analysis, oracle l; (c) PC analysis, Hallin and Liška (2007); (d) PC analysis, Bai and Ng (2002) I; (e) RSVP, $m = 20$; (f) RSVP, $m = 50$; (g) RSVP, $m = 70$; (h) RSVP–PCA

45 factors for $250 \leqslant N < 350$ and 60 factors for $N \geqslant 350$ ...'.) Because these covariates and factors are deemed the most relevant by the GTEX consortium, we refer to a data set $X_{\mathcal{T}}$ from which all $q_{\mathcal{T}}$ confounders have been removed as 'unconfounded'. However, it is possible that there is still unobserved confounding in the data sets.

For each tissue, we create a sequence of data sets by regressing out $0, 1, 2, \ldots, q_{\mathcal{T}}$ confounders. On each of these data sets, we run RSVP, PC removal with different values l of components removed. We also run the neighbourhood selection with the square-root lasso (Belloni *et al.*, 2011) on both the sample covariance matrix of the raw data set, NS, and on the covariance matrix estimated by RSVP, RSVP+NS. Two commonly used proxies for pairs of genes being co-regulated are large off-diagonal entries in the covariance or non-zero entries in the inverse covariance matrix. We therefore form for each estimated covariance matrix, a sequence of estimated co-regulation networks containing edges corresponding to the largest r entries, with r ranging from 1 to 100. In the case of NS and RSVP+NS, we vary the tuning parameter of the square-root lasso until we obtain a graph with approximately 100 edges and then form a sequence of 100 networks corresponding to the largest r entries in the estimated inverse covariance matrices, with $1 \leqslant r \leqslant 100$.

We first sought to quantify how sensitive the graphs that are returned by the various methods are to the addition of confounding. For that, for each (tissue, method, r) triple, we computed the Jaccard similarity between the edge set of a graph estimated on the unconfounded data and the graph with r edges estimated on the data set with $k \in \{0, 1, 5, 10, 20\}$ confounders removed. Fig. 11 shows the resulting Jaccard similarities averaged across the 44 tissues. Unsurprisingly the more confounders are removed, the more similar the estimated graphs are to that obtained on the unconfounded data ($k = q_{\mathcal{T}}$). However, this change for RSVP is only very slight and the method yields large similarities across different numbers of edges and k. This is an encouraging result, particularly given that a number of the confounders, such as gender and genotype data, were derived entirely from *external* data. In contrast, the performances of PC removal and NS are strongly influenced by the presence of the confounders, with the Jaccard similarity between raw and unconfounded data close to 0.

Consistently returning the same set of edges irrespectively of confounding does not imply anything about the quality of the estimates. To obtain a sense of their accuracy, we scored the graphs by using a reference data set: the gene ontology (Ashburner *et al.*, 2000). Briefly, the gene ontology is a popular database which allows the annotation of each gene by a set of *terms* classified in three categories: cellular components, molecular function and biological process. Genes that tend to perform similar functions or to interact are expected to be annotated by similar terms. By mapping each node of each graph to its gene ontology terms, one can compute a so-called enrichment statistic (Frot *et al.*, 2019a) reflecting whether the graph contains edges between related genes more often than would be expected in a random graph with a similar topology (such a graph has an expected statistic of 1). Fig. 12(a) shows the enrichment scores that were obtained in the raw data set (no confounders regressed out), averaged across all tissues. Fig. 12(b) gives the average score as a function of the number of confounders regressed out. In the online supplementary materials, the scores for each of the 44 tissues is plotted. Several comments are in order. RSVP performs well across the data sets and is the best performer on average when applied to the unconfounded data. Interestingly, as shown in the supplementary materials, there is at least one selection of l for each tissue where PC removal performs comparably with RSVP, but the optimal value of l changes from tissue to tissue. This would suggest a data-based selection for l; however, the selection criteria of Bai and Ng (2002) and Hallin and Liška (2007) both yield $l = 0$ on every tissue. The performance of the neighbourhood selection NS steadily increases as increasingly more confounders are regressed out, until it outperforms RSVP. This tends to

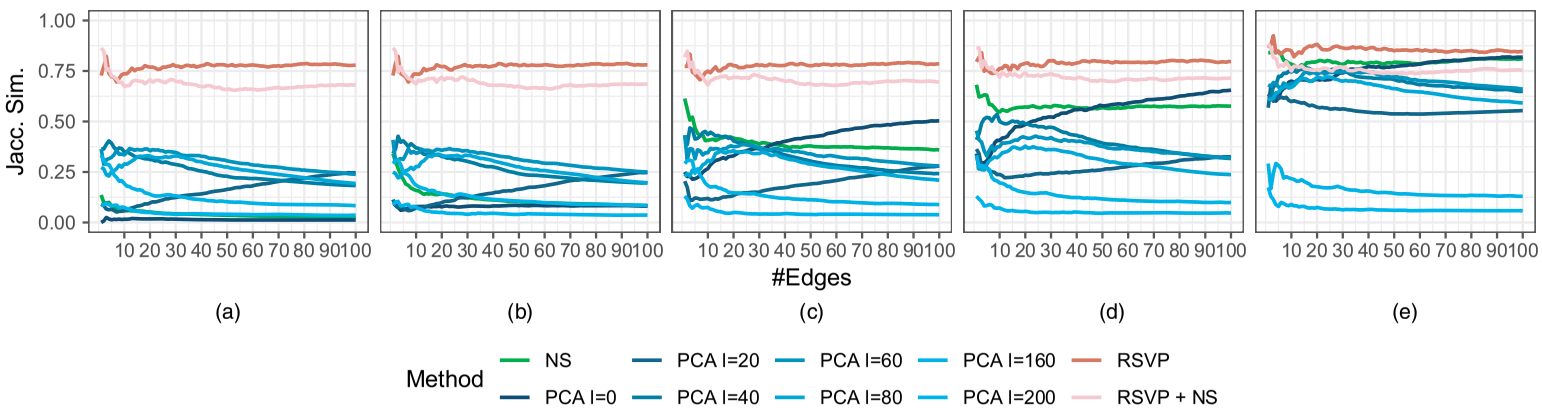

Fig. 11. Average Jaccard similarity between the edge sets of graphs estimated on the unconfounded data (with all confounders regressed out) and data from which k confounders have been removed, for $k = 0, 1, 5, 10, 20$; similarities are averaged over 44 tissues (the RSVP estimate is here seen to be the most stable with respect to removal of confounders; the RSVP estimator shows highest Jaccard similarity across all graph sizes when zero or just a few confounders are regressed out): (a) $k = 0$; (b) $k = 1$; (c) $k = 5$; (d) $k = 10$; (e) $k = 20$

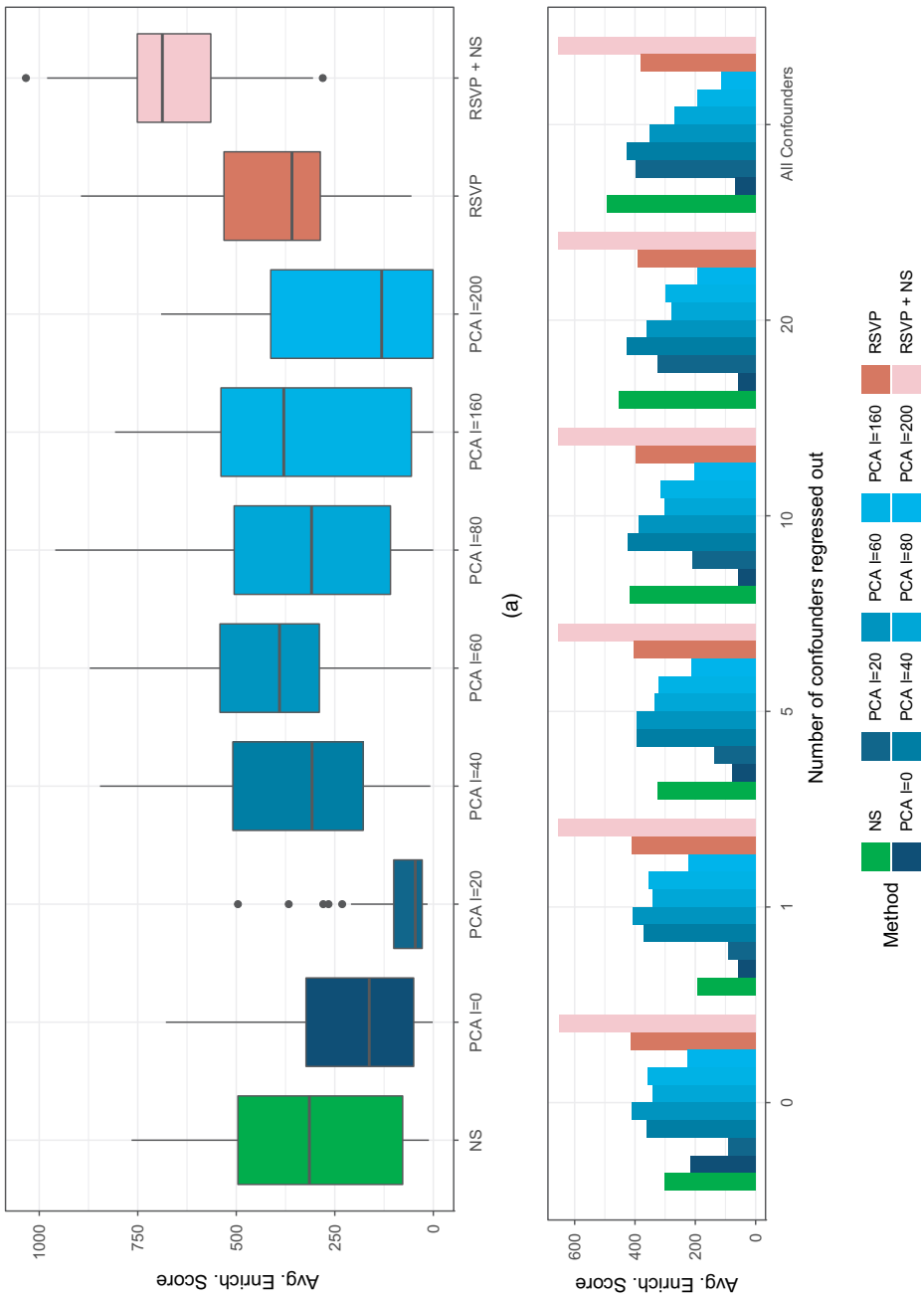

Fig. 12. (a) Area under the curve of the graph of enrichment score as a function of the number of edges, based on the raw data (we plot the distribution of the scores across the 44 tissues) and (b) average of the areas under the curve across tissues, but for data with varying numbers of confounders regressed out (the individual results for each tissue are presented in section K of the on-line supplementary material)

confirm that the raw data do indeed contain latent confounders masking true biological signal. Moreover, the fact that methods forming networks based on the estimated inverse covariances, NS and RSVP+NS, perform best on the unconfounded data sets tends to confirm that it is indeed the precision matrices which contain relevant signal when it comes to co-regulation networks.

The computational cost of performing NS is far greater than for RSVP or the PC removal approaches. We also note that the RSVP and PC removal methods may be further accelerated by using large inner product search algorithms. For example, the xyz algorithm of Thanei *et al.* (2018) can locate the large entries in the matrix product VV^{T} that forms RSVP at a fraction of the cost of performing the full matrix multiplication. On these GTEX data sets, it delivers similar performance to regular RSVP but cuts the computational cost by a factor of around 2000.

6. Discussion

In this work, we have introduced RSVP as a simple and fast method for estimating the idiosyncratic covariance Σ given data where latent factors are present. A notable aspect of the method is that all information about Σ that is contained in the spectrum of the empirical covariance matrix is thrown away. Estimation of Σ, which is permitted to have a diverging condition number, is performed by using a scaled multiple of a projection matrix whose eigenvalues are necessarily in $\{0, 1\}$. It may seem surprising at first sight that this should work at all, and the success of the method underlines the message that has emerged on the vast theory surrounding high dimensional PC analysis and covariance estimation, saying that the eigenvalues of the empirical covariance matrix $\hat{\Theta}$ are extremely noisy. By removing the variance due to these noisy eigenvalues, RSVP can cope well even in settings that are particularly challenging for PC removal approaches where the eigenvalues of the combined covariance Θ are not well separated into two groups. A drawback of RSVP is that the scale of Σ is lost, but this is of little consequence in many applications of interest and has the advantage of allowing the method to be robust to certain heavy-tailed data, for example.

Our work leaves open some questions. For example, it would be interesting to explore whether there are other estimators of the form (4) that depend on the spectrum of $\hat{\Theta}$ such that the scale of Σ is not lost, but in a sufficiently smooth way as not to have high variance even in the challenging scenarios that were mentioned above. Another interesting problem is that of controlling for latent confounding when the influence of the confounding is not linear, such as the max-linear settings (Gissibl and Klüppelberg, 2018).

Acknowledgements

Rajen Shah was supported by Engineering and Physical Sciences Research Council 'First' grant EP/R013381/1 and the Alan Turing Institute under Engineering and Physical Sciences Research Council grant EP/N510129/1.

References

Aguet, F., Brown, A. A., Castel, S., Davis, J. R., Mohammadi, P., Segre, A. V., Zappala, Z., Abell, N. S., Fresard, L., Gamazon, E. R., Gelfand, E., Gloudemans, M. J., He, Y., Hormozdiari, F., Li, X., Liu, B., Garrido Martin, D., Origen, H., Palowitch, J. J., Park, Y. S., Peterson, C. B., Quon, G., Ripke, S., Shabalin, A. A., Shimko, T. C., Strober, B. J., Sullivan, T. J., Teran, N. A., Tsang, E. K., Zhang, H., Zhou, Y.-H., Battle, A., Bustamante, C. D., Cox, N. J., Engelhardt, B. E., Eskin, E., Getz, G., Kellis, M., Li, G., MacArthur, D. G., Nobel, A. B., Sabatti, C., Wen, X., Wright, F. A., GTEX Consortium, Lappalainen, T., Ardlie, K. G., Dermitzarkis, E. T., Brown, C. D. and Montgomery, S. B. (2016) Local genetic effects on gene expression across 44 human tissues. *Nature*, **550**, 204–213.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarkis, A., Lewis, S., Matese, J. C., Richardson, J. C., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. *Nat. Genet.*, **25**, 25–29.

Bai, J. and Ng, S. (2002) Determining the number of factors in approximate factor models. *Econometrica*, **70**, 191–221.

Barigozzi, M. and Cho, H. (2018) Consistent estimation of high-dimensional factor models when the factor number is over-estimated. *Preprint arXiv:1811.00306*. London School of Economics and Political Science, London.

Belloni, A., Chernozhukov, V. and Wang, L. (2011) Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika*, **98**, 791–806.

Bickel, P. and Levina, E. (2008) Covariance regularization by thresholding. *Ann. Statist.*, **36**, 2577–2604.

Breiman, L. (1996) Bagging predictors. *Mach. Learn.*, **24**, 123–140.

Cai, T., Liu, W. and Luo, X. (2011) A constrained ℓ_1 minimization approach to sparse precision matrix estimation. *J. Am. Statist. Ass.*, **106**, 594–607.

Cai, T. T., Ren, Z. and Zhou, H. H. (2016) Estimating structured high-dimensional covariance and precision matrices: optimal rates and adaptive estimation. *Electron. J. Statist.*, **10**, 1–59.

Candès, E., Li, X., Ma, Y. and Wright, J. (2011) Robust principal component analysis? *J. Ass. Comput. Mach.*, **58**, article 11.

Ćevid, D., Bühlmann, P. and Meinshausen, N. (2018) Spectral deconfounding and perturbed sparse linear models *Preprint arXiv:1811.05352*. Eidgenössiche Technische Hochschule, Zurich.

Chandrasekaran, V., Parrilo, P. A. and Willsky, A. S. (2012) Latent variable graphical model selection via convex optimization. *Ann. Statist.*, **40**, 1935–1967.

Chandrasekaran, V., Sanghavi, S., Parrilo, P. A. and Willsky, A. S. (2011) Rank-sparsity incoherence for matrix decomposition. *SIAM J. Optimzn*, **21**, 572–596.

Chernozhukov, V., Hansen, C., Liao, Y. *et al.* (2017) A lava attack on the recovery of sums of dense and sparse signals. *Ann. Statist.*, **45**, 39–76.

Davis, C. and Kahan, W. M. (1970) The rotation of eigenvectors by a perturbation: iii. *SIAM J. Numer. Anal.*, **7**, 1–46.

Donoho, D., Gavish, M. and Johnstone, I. (2018) Optimal shrinkage of eigenvalues in the spiked covariance model. *Ann. Statist.*, **46**, 1742–1778.

Fan, J., Liao, Y. and Mincheva, M. (2013) Large covariance estimation by thresholding principal orthogonal complements (with discussion). *J. R. Statist. Soc.* B, **75**, 603–680.

Fan, J., Liu, H. and Wang, W. (2018) Large covariance estimation through elliptical factor models. *Ann. Statist.*, **46**, 1383–1414.

Friedman, J., Hastie, T. and Tibshirani, R. (2008) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**, 432–441.

Friedman, J., Hastie, T. and Tibshirani, R. (2018) glasso: graphical lasso: estimation of Gaussian graphical models. *R Package Version 1.10*. (Available from `https://CRAN.R-project.org/package=glasso`.)

Frot, B., Jostins, L. and McVean, G. (2019a) Graphical model selection for Gaussian conditional random fields in the presence of latent variables. *J. Am. Statist. Ass.*, **114**, 723–734.

Frot, B., Nandy, P. and Maathuis, M. (2019b) Robust causal structure learning with some hidden variables. *J. R. Statist. Soc.* B, **81**, 459–487.

Gagnon-Bartsch, J. A., Jacob, L. and Speed, T. P. (2013) Removing unwanted variation from high dimensional data with negative controls. *Technical Report 820*. Department of Statistics, University of California at Berkeley, Berkeley.

Gissibl, N. and Klüppelberg, C. (2018) Max-linear models on directed acyclic graphs. *Bernoulli*, **24**, no. 4A, 2693–2720.

Haavelmo, T. (1944) The probability approach in econometrics. *Econometrica*, **12**, 1–115.

Hallin, M. and Liška, R. (2007) Determining the number of factors in the general dynamic factor model. *J. Am. Statist. Ass.*, **102**, 603–617.

Harris, N. and Drton, M. (2013) PC algorithm for nonparanormal graphical models. *J. Mach. Learn. Res.*, **14**, 3365–3383.

Heinze-Deml, C., Maathuis, M. and Meinshausen, N. (2018) Causal structure learning. *A. Rev. Statist. Appl.*, **5**, 371–391.

Jia, J. and Rohe, K. (2015) Preconditioning the lasso for sign consistency. *Electron. J. Statist.*, **9**, 1150–1172.

Kalisch, P. and Bühlmann, P. (2007) Estimating high-dimensional directed acyclic graphs with the PC-algorithm. *J. Mach. Learn. Res.*, **8**, 613–636.

Klochkov, Y. and Zhivotovskiy, N. (2018) Uniform Hanson-Wright type concentration inequalities for unbounded entries via the entropy method. *Preprint arXiv:1812.03548*. Humboldt University, Berlin.

Lauritzen, S. (1996) *Graphical Models*. Oxford: Oxford University Press.

Ledoit, O. and Wolf, M. (2004) A well-conditioned estimator for large-dimensional covariance matrices. *J. Multiv. Anal.*, **88**, 365–411.

Leek, J. T. and Storey, J. D. (2007) Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLOS Genet.*, **3** article e161.
Meek, C. (1995) Strong completeness and faithfulness in Bayesian networks. In *Uncertainty in Artificial Intelligence* (eds P. Besnard and S. Hanks), pp. 411–418. San Francisco: Morgan Kaufmann.
Meinshausen, N. and Bühlmann, P. (2006) High dimensional graphs and variable selection with the Lasso. *Ann. Statist.*, **34**, 1436–1462.
Menchero, J., Morozov, A. and Shepard, P. (2010) Global equity risk modeling. In *Handbook of Portfolio Construction*, pp. 439–480. Berlin: Springer.
Pearl, J. (2009) *Causality*. Cambridge: Cambridge University Press.
Ren, Z., Sun, T., Zhang, C.-H. and Zhou, H. H. (2015) Asymptotic normality and optimalities in estimation of large Gaussian graphical models. *Ann. Statist.*, **43**, 991–1026.
Robins, J. (1986) A new approach to causal inference in mortality studies with a sustained exposure period: application to control of the healthy worker survivor effect. *Math. Modllng*, **7**, 1393–1512.
Rohe, K. (2015) Preconditioning for classical relationships: a note relating ridge regression and OLS *p*-values to preconditioned sparse penalized regression. *Stat*, **4**, 157–166.
Spirtes, P., Glymour, C. and Scheines, R. (2000) *Causation, Prediction, and Search*, 2nd edn. Cambridge: MIT Press.
Stegle, O., Parts, L., Piipari, M., Winn, J. and Durbin, R. (2012) Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. *Nat. Protcls*, **7**, 500–507.
Thanei, G.-A., Meinshausen, N. and Shah, R. D. (2018) The xyz algorithm for fast interaction search in high-dimensional data. *J. Mach. Learn. Res.*, **19**, 1343–1384.
Wang, X. and Leng, C. (2015) High dimensional ordinary least squares projection for screening variables. *J. R. Statist. Soc.* B, **78**, 589–611.
Yuan, M. (2010) High dimensional inverse covariance matrix estimation via linear programming. *J. Mach. Learn. Res.*, **11**, 2261–2286.
Yuan, M. and Lin, Y. (2007) Model selection and estimation in the Gaussian graphical model. *Biometrika*, **94**, 19–35.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Supplementary material for "Right singular vector projection graphs: fast high dimensional covariance matrix estimation under latent confounding"'.