Notes on multivariate modeling of gene expression data

Wessel N. van Wieringen[1,2,], **Carel F.W. Peeters**[1]

[1] Department of Epidemiology and Biostatistics, VU University Medical Center

P.O. Box 7075, 1007 MB Amsterdam, The Netherlands

[2] Department of Mathematics, VU University Amsterdam

De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

Email: w.vanwieringen@vumc.nl

Abstract

Notes on the application of time series machinery to the multivariate analysis of gene expression data from a time-course experiment. The intend is to include time-varying covariates, i.e. data from other molecular levels, in the model and analysis. These notes are not meant for publication.

Acknowledgment: Mark A. van de Wiel provided feedback on earlier versions of this document that lead to this improved version.

Disclaimer: This document is incomplete and does most likely contain errors, which are the full responsibility of the authors.

Contents

1 Some necessary biology

1.1 Molecular biology

All organisms are made of cells. A cell is the smallest independently living unit. The cell harbors many molecular entities, each fulfilling a specific role in the functioning of the cell. The molecular entities come in many different types (e.g., DNA, mRNA, et cetera) here referred to as molecular levels. Some of the key players involved in the regulatory mechanism of the cell are introduced.

The DNA molecule is a double-stranded polymer composed of four nucleotides. The nucleus of human cells contains DNA molecules in the form of chromosomes, 22 autosomal and one sex chromosome pair. Hence, any part of a DNA molecule belonging to one of the autosomal chromosomes is present twice in a normal human cell. Together the haploid (single) set of chromosomes make up the complete genetic constitution of an organism, the genome. The DNA molecules carry, in the sequence of nucleotides, the information necessary for the functioning of cells, which is encoded in molecular units called genes. A gene is the basic physical unit of heredity. A gene is said to be expressed if the product that it encodes for has been formed. The central dogma of molecular biology describes the information transfer process that leads from the information encoded in DNA (genes) to the proteins of the cell. Three primary steps in this information processes are discerned: replication, transcription, and translation. Replication refers to the process of DNA duplication. Transcription is the process in which messenger RNA (mRNA) molecules are synthesized (copying of genetic information) from the DNA in the cell nucleus. Much like DNA, an mRNA molecule is a chain of nucleotides, but single stranded and consisting of a different set of nucleotides. The sequence of nucleotides, arranged into trios (codons), of an mRNA encodes for a sequence of amino acids (each codon corresponds to a specific amino acid) that form a protein. After transcription, the mRNA molecules are transported from the nucleus to the cytoplasm, where they are converted to proteins (translation). A protein is a molecule consisting of a sequence of amino acids, whose order is determined by the base sequence of nucleotides in the mRNA (gene) encoding for the protein. Proteins are important molecules for the functioning of the cell as they participate in many of its processes. Proteins are not considered in the remainder.

1.2 Cancer

Cancer is a disease in which cells exhibit unproliferated growth and can invade nearby tissues (so-called cancer cells). Cancer is a genetic disease, often caused by abnormalities in the genetic material of cancer cells. An example of such abnormalities are chromosomal aberrations, a structural abnormality in the chromosomes. This can be an abnormal DNA copy number, in which case the number of copies of a particular genomic segment deviates from the usual two (assuming the segment maps to an autosomal chromosome). Such chromosomal aberrations are a key event in the development and progression of cancer Lengauer *et al.* (1998). Chromosomal aberrations like an increase (or decrease) of the DNA copy number of a particular genomic segment are, through the central dogma of biology, likely to result in increased (or decreased) mRNA transcription levels of genes in the segment.

Among others aforementioned abnormalities affect so-called cancer genes. Cancer genes have the ability to direct malignant cell growth. Two types of cancer genes are distinguished: proto-onco genes and tumor-suppressor genes Vogelstein and Kinzler (2004). Proto-onco genes often have a gain (more than two copies) in DNA copy number, which causes increased proliferation. Tumor-suppressor genes often have a loss (less than two copies) in DNA copy number, which causes impairment of growth inhibition. For more on cancer refer to Weinberg (2006).

1.3 Measurement devices

High-throughput techniques are used to quantify the type and amount of information on the different molecular levels of the cell. For instance, DNA copy numbers can be measured by array CGH Pinkel and Albertson (2005), and mRNA levels by gene expression microarrays Nguyen *et al.* (2002). Both are a type of microarray. Conceptually, a microarray is a large collection of gauges. Each gauge measures a characteristic (e.g., DNA copy number, gene expression) of a predefined molecular entity (e.g. genomic segment, transcript). The result of measuring a sample's gene expression by a microarray is a vector of (say) 10000 values on a continuous scale. These values are preprocessed to remove experimental artifacts and arrive at the quantity of interest (refer to Nguyen *et al.* (2002) for more details on the preprocessing of gene expression data). The preprocessed values are believed to be representative (proportionally) to the expression levels of the corresponding genes. The vector of preprocessed values is referred to as the sample's gene expression profile.

Using array CGH, one obtains a data vector of similar size (although current platforms produce much larger vectors). After preprocessing, the values now represent the *relative* DNA copy number (on a continuous scale). Relative, for the DNA copy number of the sample of interest is measured in comparison to a reference sample that is believed to have a DNA copy number of two for the autosomale chromosomes. Often the array CGH data undergo further processing steps: segmentation and calling. Segmentation divides the genome into non-overlapping segments separated by breakpoints. These breakpoints indicate a change in DNA copy number, and, hence, DNA copy number is constant in between breakpoints. Segmentation may be viewed as a piece-wise constant smoothing of the data. As a final and last pre-processing step, referred to as calling, the DNA copy number of each segment is determined. At present calling algorithms cannot determine whether there are, say, three or four copies present. They can however detect deviations from the normal copy number, and classify each segment as either 'normal', 'loss', 'gain' or 'amplification': 'normal' if there are two copies of the chromosomal segment present, 'loss' (also named deletion) if at least one copy is lost, 'gain' if at least one additional copy is present, and 'amplification' if there are high level, say > 5, copy numbers (for simplicity we ignore the existence of polyploid genomes). These labels are referred to as calls.

Add microRNAs. Sequencing data. Pathways.

2 Regression

Introduction, simple linear regression.

$$\|\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}\|_2^2$$

$$\mathbf{X}_{i*}$$

$$\begin{aligned} S(\boldsymbol{\beta}) &= S(\beta_0, \beta_1) \\ &= \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \\ &= \sum_{i=1}^{n} (Y_i - E(Y_i))^2 \end{aligned}$$

$$\hat{Y}_i = E(Y_i) = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

$$\hat{\sigma}^2 = RSS/(n-2)$$

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_i + \varepsilon_i \\ E(\varepsilon_i) &= 0 \\ \mathrm{Cov}(\varepsilon_{i_1}, \varepsilon_{i_2}) &= \begin{cases} \sigma^2 & \text{if} \quad i_1 = i_2 \\ 0 & \text{if} \quad i_1 \neq i_2 \end{cases} \end{aligned}$$

2.1 Experiment and

2.2 Model

The multiple linear regression model is

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

with $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ and ε_i independent, i.e.:

$$\mathrm{Cov}(\varepsilon_{i_1}, \varepsilon_{i_2}) = \begin{cases} \sigma^2 & \text{if} \quad i_1 = i_2 \\ 0 & \text{if} \quad i_1 \neq i_2 \end{cases}$$

The model thus contains the following unknown parameters: $\beta_0, \beta_1, \ldots, \beta_{p-1}$ and σ^2.

$$E(Y_i) = \mathbf{X}_{i*}\,\boldsymbol{\beta}$$

The linear regression model can be written more conveniently in matrix notation as:

$$Y_i \;=\; \mathbf{X}_{i,*}\,\boldsymbol{\beta} + \varepsilon_i.$$

Or, even more condensed:

$$\mathbf{Y} \;=\; \mathbf{X}\,\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where

$$\begin{aligned}
\mathbf{Y} &= (Y_1, Y_2, \ldots, Y_n)^T, \\
\boldsymbol{\beta} &= (\beta_0, \beta_1, \ldots, \beta_{p-1})^T, \\
\boldsymbol{\varepsilon} &= (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T,
\end{aligned}$$

and

$$\mathbf{X} \;=\; \begin{pmatrix}
1 & X_{11} & X_{12} & \ldots & X_{1,p-1} \\
1 & X_{21} & X_{22} & \ldots & X_{2,p-1} \\
\vdots & \vdots & \vdots & \vdots & \vdots \\
1 & X_{n1} & X_{n2} & \ldots & X_{n,p-1}
\end{pmatrix}.$$

The distributional assumptions in matrix formulation become

$$\begin{aligned}
\mathrm{Cov}(\boldsymbol{\varepsilon}, \boldsymbol{\varepsilon}) &= \boldsymbol{\Sigma} = \sigma^2 \mathbf{I}_{n\times n} \\
&= \begin{pmatrix}
\mathrm{Cov}(\varepsilon_1, \varepsilon_1) & \mathrm{Cov}(\varepsilon_1, \varepsilon_2) & \ldots & \mathrm{Cov}(\varepsilon_1, \varepsilon_n) \\
\mathrm{Cov}(\varepsilon_2, \varepsilon_1) & \mathrm{Cov}(\varepsilon_2, \varepsilon_2) & \ldots & \mathrm{Cov}(\varepsilon_2, \varepsilon_n) \\
\vdots & \vdots & \vdots & \vdots \\
\mathrm{Cov}(\varepsilon_n, \varepsilon_1) & \mathrm{Cov}(\varepsilon_n, \varepsilon_2) & \ldots & \mathrm{Cov}(\varepsilon_n, \varepsilon_n)
\end{pmatrix} \\
&= \begin{pmatrix}
\sigma^2 & 0 & \ldots & 0 \\
0 & \sigma^2 & \ldots & 0 \\
\vdots & \vdots & \vdots & \vdots \\
0 & 0 & \ldots & \sigma^2
\end{pmatrix}
\end{aligned}$$

$$\begin{aligned}
E(\boldsymbol{\varepsilon}) &= (E(\varepsilon_1), E(\varepsilon_2), \ldots, E(\varepsilon_n))^T \\
&= (0, 0, \ldots, 0)^T \;=\; \mathbf{0}
\end{aligned}$$

$$\begin{aligned}
E(\mathbf{Y}) &= \begin{pmatrix}
E(Y_1) \\ E(Y_2) \\ \vdots \\ E(Y_n)
\end{pmatrix} \\
&= \begin{pmatrix}
\beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \ldots + \beta_{p-1} X_{1,p-1} \\
\beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \ldots + \beta_{p-1} X_{2,p-1} \\
\vdots \\
\beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \ldots + \beta_{p-1} X_{n,p-1}
\end{pmatrix} \\
&= \mathbf{X}\,\boldsymbol{\beta}
\end{aligned}$$

2.3 Estimation

The parameters of the regression model, $\boldsymbol{\beta}$ and σ^2 are estimated by means of likelihood maximization. Recall that $Y_i \sim \mathcal{N}(\mathbf{X}_{i,*}\,\boldsymbol{\beta}, \sigma^2)$ with corresponding density:

$$f_{Y_i}(y_i) \;=\; \frac{1}{\sqrt{2\,\pi}\,\sigma}\,\exp[-(y_i - \mathbf{X}_{i*}\,\boldsymbol{\beta})/2\sigma^2].$$

The likelihood thus is:

$$\begin{aligned} L(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) \;&=\; f_{\mathbf{Y}}(\mathbf{y}) \;=\; \prod_{i=1}^{n} f_{Y_i}(y_i) \\ &=\; \prod_{i=1}^{n} \frac{1}{\sqrt{2\,\pi}\,\sigma}\,\exp[-(y_i - \mathbf{X}_{i,*}\,\boldsymbol{\beta})/2\sigma^2]. \end{aligned}$$

Because of the concavity of the logarithm, the maximization of the likelihood coincides with the maximum of the logarithm of the likelihood (called the log-likelihood). Hence, to obtain maximum likelihood estimates of the parameter it is equivalent to find the maximum of the log-likelihood. The log-likelihood is:

$$\begin{aligned} \mathcal{L}(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) \;&=\; \log[L(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2)] \;=\; \log\Big[\prod_{i=1}^{n} f_{Y_i}(y_i)\Big] \;=\; \sum_{i=1}^{n} \log[f_{Y_i}(y_i)] \\ &=\; \sum_{i=1}^{n} [-\log(\sqrt{2\,\pi}\,\sigma) - (y_i - \mathbf{X}_{i*}\,\boldsymbol{\beta})^2/2\sigma^2] \\ &=\; -n\,\log(\sqrt{2\,\pi}\,\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mathbf{X}_{i,*}\,\boldsymbol{\beta})^2. \end{aligned}$$

After noting that:

$$\sum_{i=1}^{n}(y_i - \mathbf{X}_{i,*}\,\boldsymbol{\beta})^2 \;=\; \|\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}\|_2^2 \;=\; (\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta})^T\,(\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}),$$

the log-likelihood can be written as:

$$\mathcal{L}(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) \;=\; -n\,\log(\sqrt{2\,\pi}\,\sigma) - \frac{1}{2\sigma^2}\,\|\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}\|_2^2.$$

In order to find the maximum of the log-likelihood, take its derivate with respect to $\boldsymbol{\beta}$:

$$\begin{aligned} \frac{\partial}{\partial\,\beta}\mathcal{L}(\mathbf{Y}, \mathbf{X}; \boldsymbol{\beta}, \sigma^2) \;&=\; -\frac{1}{2\sigma^2}\,\frac{\partial}{\partial\,\beta}\|\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}\|_2^2 \\ &=\; \frac{1}{\sigma^2}\,\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}). \end{aligned}$$

Equate this derivate to zero gives the estimating equation for $\boldsymbol{\beta}$:

$$\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\,\boldsymbol{\beta}) \;=\; \mathbf{0}.$$

Expand and reshuffle this equation to:

$$\mathbf{X}^T\mathbf{X}\,\boldsymbol{\beta} \;=\; \mathbf{X}^T\mathbf{Y}$$

Pre-multiplication of both sides by $(\mathbf{X}^T\mathbf{X})^{-1}$ now yields the ML estimator of the regression parameter:

$$\hat{\boldsymbol{\beta}} \;=\; (\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{Y}.$$

Along the same lines one obtains the ML estimator of the residual variance. Take the partial derivative of the log-likelihood with respect to σ^2:

$$\frac{\partial}{\partial\,\sigma}\mathcal{L}(\mathbf{Y},\mathbf{X};\boldsymbol{\beta},\sigma^2) \;=\; -\frac{n}{\sigma}+\frac{1}{\sigma^3}\|\mathbf{Y}-\mathbf{X}\,\boldsymbol{\beta}\|_2^2.$$

Equate the right-hand side to zero and solve for σ^2:

$$\hat{\sigma}^2 \;=\; \frac{1}{n}\|\mathbf{Y}-\mathbf{X}\,\boldsymbol{\beta}\|_2^2.$$

In this expression $\boldsymbol{\beta}$ is unknown and in practice the ML estimate of $\boldsymbol{\beta}$ is plugged-in.

We now focus on the properties of the derived ML estimators. The expectation of the ML estimator of the regression parameter $\boldsymbol{\beta}$ is:

$$\begin{aligned} E(\hat{\boldsymbol{\beta}}) &= E[(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\mathbf{Y}] \\ &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}E[\mathbf{Y}] \\ &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\mathbf{X}\,\boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned}$$

Hence, the ML estimator of the regression coefficients is unbiased.

The variance of the ML estimator of $\boldsymbol{\beta}$ is:

$$\begin{aligned} \boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}\hat{\boldsymbol{\beta}}} &= \mathrm{Cov}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}) \\ &= E\{[\hat{\boldsymbol{\beta}}-E(\hat{\boldsymbol{\beta}})][\hat{\boldsymbol{\beta}}-E(\hat{\boldsymbol{\beta}})]^{\mathrm{T}}\} \\ &= E\{[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\mathbf{Y}-\boldsymbol{\beta}]\,[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\mathbf{Y}-\boldsymbol{\beta}]^{\mathrm{T}}\} \\ &= E\{[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\mathbf{Y}]\,[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\mathbf{Y}]^{\mathrm{T}}\}-\boldsymbol{\beta}\,\boldsymbol{\beta}^{\mathrm{T}} \\ &= E\{(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\mathbf{Y}\,\mathbf{Y}^{\mathrm{T}}\,\mathbf{X}\,(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\}-\boldsymbol{\beta}\,\boldsymbol{\beta}^{\mathrm{T}} \\ &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\,E\{\mathbf{Y}\,\mathbf{Y}^{\mathrm{T}}\}\,\mathbf{X}\,(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}-\boldsymbol{\beta}\,\boldsymbol{\beta}^{\mathrm{T}} \\ &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\,\mathbf{X}^{\mathrm{T}}\,\{\mathbf{X}\,\boldsymbol{\beta}\,\boldsymbol{\beta}^{\mathrm{T}}\,\mathbf{X}^{\mathrm{T}}+\boldsymbol{\Sigma}\}\,\mathbf{X}\,(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}-\boldsymbol{\beta}\,\boldsymbol{\beta}^{\mathrm{T}} \\ &= \boldsymbol{\beta}\,\boldsymbol{\beta}^{\mathrm{T}}+\sigma^2\,(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}-\boldsymbol{\beta}\,\boldsymbol{\beta}^{\mathrm{T}} \\ &= \sigma^2\,(\mathbf{X}^T\mathbf{X})^{-1}, \end{aligned}$$

in which we have used that

$$\begin{aligned} E(\mathbf{Y}\mathbf{Y}^{\mathrm{T}}) &= \mathbf{X}\,\boldsymbol{\beta}\,\boldsymbol{\beta}^{\mathrm{T}}\,\mathbf{X}^{\mathrm{T}}+\boldsymbol{\Sigma} \\ &= \mathbf{X}\,\boldsymbol{\beta}\,\boldsymbol{\beta}^{\mathrm{T}}\,\mathbf{X}^{\mathrm{T}}+\sigma^2\,\mathbf{I}_{n\times n}. \end{aligned}$$

From $\mathrm{Cov}(\hat{\boldsymbol{\beta}},\hat{\boldsymbol{\beta}}) = \sigma^2\,(\mathbf{X}^T\mathbf{X})^{-1}$ we obtain a natural estimate of the standard error of the estimate of the j-th regression coefficient: $\hat{\sigma}_{\hat{\beta}_j} = \hat{\sigma}\sqrt{[(\mathbf{X}^T\mathbf{X})^{-1}]_{jj}}$.

Insert the expectation of the variance estimate. Is biased.

Estimation of residuals:

$$\begin{aligned}\hat{\boldsymbol{\varepsilon}} &= \mathbf{Y} - \hat{\mathbf{Y}} \\ &= \mathbf{Y} - \mathbf{X}\,\hat{\boldsymbol{\beta}} \\ &= \mathbf{Y} - \mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= [\mathbf{I} - \mathbf{X}\,(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T]\,\mathbf{Y}\end{aligned}$$

Thus, the residuals are a projection of $\mathbf{Y}$ onto the orthogonal complement of space spanned by the columns of $\mathbf{X}$.

To be used in diagnostics.

2.4 Testing

For each parameter we test the null hypothesis:

$$H_0 : \beta_j = 0.$$

To evaluate this hypothesis we note that:

$$\frac{\hat{\beta}_j - \beta}{\hat{\sigma}_{\hat{\beta}_j}} \quad \sim \quad t_{n-p}$$

where

2.5 Coefficient of determination

The coefficient of determination is defined as:

$$R^2(\mathbf{Y},\,\mathbf{X}) \quad = \quad [\rho(\mathbf{Y},\,\hat{\mathbf{Y}})]^2 \quad = \quad [\rho(\mathbf{Y},\,\mathbf{X}\hat{\boldsymbol{\beta}})]^2,$$

the squared correlation coefficient between $\mathbf{Y}$ and the columns of $\mathbf{X}$. Often, $R^2(\mathbf{Y},\,\mathbf{X})$ is simple referred to as R^2. From the definition is clear that $R^2 \in [0, 1]$ as $\rho \in [-1, 1]$.

An alternative interpretation of the coefficient of determination R^2 comes from the sum of squares of the observations:

$$SYY \quad = \quad \sum_{i=1}^{n}(Y_i - \bar{Y})^2 \quad = \quad \|\mathbf{Y} - \bar{\mathbf{Y}}\|_2^2.$$

We may then write:

$$R^2 \quad = \quad \frac{SYY - RSS}{SYY} \quad = \quad \frac{SYY/(n-1) - RSS/(n-1)}{SYY/(n-1)} \quad = \quad \frac{s_Y^2 - s_{\hat{\varepsilon}}^2}{s_Y^2},$$

where

$$\begin{aligned}s_{\hat{\varepsilon}}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(\varepsilon_i - \bar{\hat{\varepsilon}})^2 \quad = \quad \frac{1}{n-1}\sum_{i=1}^{n}\varepsilon_i^2 \\ &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2 \quad = \quad RSS/(n-1).\end{aligned}$$

The "percentage of explained variation" in $\mathbf{Y}$ by $\mathbf{X}$.

$$RSS \;=\; \|\mathbf{Y} - \hat{\mathbf{Y}}\|^2 \;=\; \|\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}\|^2$$

$$\hat{\sigma}^2 \;=\; s^2 \;=\; RSS/(n-p)$$

2.6 Illustration

2.7 Exercises

2.8 Answers

3 Ridge regression

3.1 Introduction

Collinearity in regression analysis refers to the event of two (or multiple) covariates being highly linearly related. When fitting a linear regression model to data, collinearity may cause large standard error of estimates.

The case of two (or multiple) covariates being perfectly linearly dependent is referred as super-collinearity.

Example 1. *Super-collinearity*
Consider the design matrix:

$$\mathbf{X} = \begin{pmatrix} 1 & -1 & 2 \\ 1 & 0 & 1 \\ 1 & 2 & -1 \\ 1 & 1 & 0 \end{pmatrix}$$

The columns of $\mathbf{X}$ are linearly dependent: the first column is the row-wise sum of the other two columns.

As a consequence of super-collinearity of a design matrix $\mathbf{X}$, the matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is singular. A square matrix that does not have an inverse is called *singular*. A matrix $\mathbf{A}$ is singular if and only if its determinant is zero: $\det(A) == 0$.

Example 2. *Singularity*
Consider the matrix $\mathbf{A}$ given by:

$$\mathbf{A} = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}$$

Clearly, $\det(A) = a_{11}a_{22} - a_{12}a_{21} = 1 \times 4 - 2 \times 2 = 0$. Hence, $\mathbf{A}$ is singular, and its inverse is undefined.

As $\det(\mathbf{A})$ is equal to the product of the eigenvalues λ_j of $\mathbf{A}$, the matrix $\mathbf{A}$ is singular if any of the eigenvalues of $\mathbf{A}$ is zero. To see this, consider the spectral decomposition of $\mathbf{A}$:

$$\mathbf{A} = \sum_{j=1}^{p} \lambda_j \, \mathbf{v} \, \mathbf{v}_j,$$

where $\mathbf{v}_j$ is the eigenvector corresponding to λ_j. The inverse of $\mathbf{A}$ is then:

$$\mathbf{A}^{-1} = \sum_{j=1}^{p} \lambda_j^{-1} \, \mathbf{v} \, \mathbf{v}_j.$$

The right-hand side is undefined if $\lambda_j = 0$ for any j.

Example 3. *Singularity*
Revisit Example 2. Matrix $\mathbf{A}$ has eigenvalues $\lambda_1 = 5$ and $\lambda_2 = 0$. According to the spectral decomposition, the inverse of $\mathbf{A}$ is:

$$\mathbf{A}^{-1} = \frac{1}{5} \, \mathbf{v}_1 \, \mathbf{v}_1^{\mathrm{T}} + \frac{1}{0} \, \mathbf{v}_2 \, \mathbf{v}_2^{\mathrm{T}}.$$

This expression is undefined as we divide by zero in the second summand on the right-hand side.

Super-collinearity causes $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ to be singular. So? Recall the estimator of the regression coefficients (and its variance):

$$\hat{\boldsymbol{\beta}} \;=\; (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \tag{1}$$

These are only defined if $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ exits.Hence, supercollinearity: regression coefficients cannot be estimated.

Overfitting.

In case of a singular matrix $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ its inverse is not $(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ defined, and, consequently, the OLS estimator of the regression coefficients $\hat{\boldsymbol{\beta}} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$.

An ad-hoc solution to the singularity of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is to add the term $\lambda\mathbf{I}_{p\times p}$, leading to an 'ad-hoc OLS' estimator:

$$\hat{\boldsymbol{\beta}}(\lambda) \;=\; (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_{p\times p})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}.$$

This estimator is called the *ridge estimator*.

3.2 Expectation

The expectation of the ridge estimator is:

$$\begin{aligned}
E\big[\hat{\boldsymbol{\beta}}(\lambda)\big] &= E\big[(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{p\times p})^{-1}\mathbf{X}^T\mathbf{Y}\big] \\
&= E\big\{[\mathbf{I}_{p\times p} + \lambda\,(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\,\mathbf{X}^T\mathbf{Y}\big\} \\
&= E\big\{[\mathbf{I}_{p\times p} + \lambda\,(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}\,\hat{\boldsymbol{\beta}}\big\} \\
&= [\mathbf{I}_{p\times p} + \lambda\,(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}\,E\big(\hat{\boldsymbol{\beta}}\big) \\
&= [\mathbf{I}_{p\times p} + \lambda\,(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}\,\boldsymbol{\beta}.
\end{aligned}$$

Clearly, $E\big[\hat{\boldsymbol{\beta}}(\lambda)\big] \neq \boldsymbol{\beta}$. Hence, the ridge estimator is biased. The biasedness of the ridge estimator is biased could also have been concluded from the Gauss-Markov Theorem **expand**.

The expectation of the ridge estimator vanishes as λ tends to infinity:

$$\lim_{\lambda\to\infty} E\big[\hat{\boldsymbol{\beta}}(\lambda)\big] \;=\; \lim_{\lambda\to\infty}[\mathbf{I}_{p\times p} + \lambda\,(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}\,\boldsymbol{\beta} \;=\; \mathbf{0}_{p\times 1}.$$

Hence, all regression coefficients are shrunken towards zero as the penalty parameter increases.

Example 4. *Orthonormal design matrix*
Consider an orthonormal design matrix $\mathbf{X}$, i.e.:

$$\mathbf{X}^T\mathbf{X} \;=\; \mathbf{I}_{p\times p} \;=\; (\mathbf{X}^T\mathbf{X})^{-1}.$$

An example of an orthonormal design matrix would be:

$$\mathbf{X} \;=\; \frac{1}{2}\begin{pmatrix} -1 & -1 \\ -1 & 1 \\ 1 & -1 \\ 1 & 1 \end{pmatrix}.$$

This design matrix is orthonormal as $\mathbf{X}^{\mathrm{T}}\mathbf{X} = \mathbf{I}_{2\times 2}$, which is easily verified:

$$\mathbf{X}^{\mathrm{T}}\mathbf{X} \;=\; \frac{1}{4}\begin{pmatrix} -1 & -1 & 1 & 1 \\ -1 & 1 & -1 & 1 \end{pmatrix}\begin{pmatrix} -1 & -1 \\ -1 & 1 \\ 1 & -1 \\ 1 & 1 \end{pmatrix} \;=\; \frac{1}{4}\begin{pmatrix} 4 & 0 \\ 0 & 4 \end{pmatrix}.$$

In case of an orthonormal design matrix the relation between the OLS and ridge estimator is:

$$\begin{aligned} \hat{\boldsymbol{\beta}}(\lambda) &= (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (\mathbf{I} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (1+\lambda)^{-1}\mathbf{I}\mathbf{X}^T\mathbf{Y} \\ &= (1+\lambda)^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \\ &= (1+\lambda)^{-1}\hat{\boldsymbol{\beta}}. \end{aligned}$$

Hence, the ridge estimator scales the OLS estimator by a factor. When taking the expectation on both sides, it is evident that the ridge estimator converges to zero as $\lambda \to \infty$.

3.3 Variance

In order to derive the variance of the ridge estimator define:

$$\mathbf{W}_\lambda \;=\; [\mathbf{I} + \lambda(\mathbf{X}^T\mathbf{X})^{-1}]^{-1}.$$

Using $\mathbf{W}_\lambda$ the ridge estimator $\hat{\boldsymbol{\beta}}(\lambda)$ can be expressed as $\mathbf{W}_\lambda\hat{\boldsymbol{\beta}}$ for:

$$\begin{aligned} \mathbf{W}_\lambda\hat{\boldsymbol{\beta}} &= \mathbf{W}_\lambda(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ &= \{(\mathbf{X}^T\mathbf{X})^{-1}[\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{p\times p}]\}^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ &= [\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{p\times p}]^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ &= [\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}_{p\times p}]^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ &= \hat{\boldsymbol{\beta}}(\lambda). \end{aligned}$$

It is now easily seen that:

$$\begin{aligned} \mathrm{Var}[\hat{\boldsymbol{\beta}}(\lambda)] &= \mathrm{Var}[\mathbf{W}_\lambda\hat{\beta}] \\ &= \mathbf{W}_\lambda\mathrm{Var}[\hat{\beta}]\mathbf{W}_\lambda^{\mathrm{T}} \\ &= \sigma^2\mathbf{W}_\lambda(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{W}_\lambda^{\mathrm{T}}, \end{aligned}$$

in which we have used $\mathrm{Var}(\mathbf{AY}) = \mathbf{A}\mathrm{Var}(\mathbf{Y})\mathbf{A}^{\mathrm{T}}$ for a non-random $\mathbf{A}$, the fact that $\mathbf{W}_\lambda$ is non-random, and $\mathrm{Var}[\hat{\beta}] = \sigma^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$.

As the expectation, the variance of the ridge estimator vanishes as λ tends to infinity:

$$\lim_{\lambda\to\infty}\mathrm{Var}\big[\hat{\boldsymbol{\beta}}(\lambda)\big] \;=\; \sigma^2\mathbf{W}_\lambda(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{W}_\lambda^{\mathrm{T}} \;=\; \mathbf{0}_{p\times p}.$$

Hence, the variance of the ridge regression coefficient estimates decreases towards zero as the penalty parameter becomes large.

With an explicit expression of the variance of the ridge estimator at hand, we can compare it to that of the OLS estimator:

$$\begin{aligned} \mathrm{Var}[\hat{\boldsymbol{\beta}}] - \mathrm{Var}[\hat{\boldsymbol{\beta}}(\lambda)] &= \sigma^2[(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} - \mathbf{W}_\lambda(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{W}_\lambda^{\mathrm{T}}] \\ &= \sigma^2\mathbf{W}_\lambda\{[\mathbf{I} + \lambda(\mathbf{X}^T\mathbf{X})^{-1}](\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}[\mathbf{I} + \lambda(\mathbf{X}^T\mathbf{X})^{-1}]^{\mathrm{T}} - (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\}\mathbf{W}_\lambda^{\mathrm{T}} \\ &= \sigma^2\mathbf{W}_\lambda[2\,\lambda\,(\mathbf{X}^T\mathbf{X})^{-2} + \lambda^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-3}]\mathbf{W}_\lambda^{\mathrm{T}}. \end{aligned}$$

The difference is non-negative definite as each component in the matrix product is non-negative definite. Hence,

$$\text{Var}[\hat{\boldsymbol{\beta}}] \quad \succeq \quad \text{Var}[\hat{\boldsymbol{\beta}}(\lambda)].$$

In words, the variance of the OLS estimator is larger than that of the ridge estimator (in the sense that their difference is non-negative definite). **Visualize this!**

Example 5. *Orthonormal design matrix*
Assume the design matrix $\mathbf{X}$ is orthonormal. Then, $\text{Var}[\hat{\boldsymbol{\beta}}] = \sigma^2 \mathbf{I}_{p\times p}$ and

$$\begin{aligned}
\text{Var}[\hat{\boldsymbol{\beta}}(\lambda)] &= \sigma^2 \mathbf{W}_\lambda (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{W}_\lambda^{\mathrm{T}} \\
&= \sigma^2 [\mathbf{I}_{p\times p} + \lambda \mathbf{I}_{p\times p}]^{-1} \mathbf{I}_{p\times p} \{[\mathbf{I} + \lambda \mathbf{I}_{p\times p}]^{-1}\}^{\mathrm{T}} \\
&= \sigma^2 (1+\lambda)^{-2} \mathbf{I}_{p\times p}.
\end{aligned}$$

As the penalty parameter λ is non-negative the former exceeds the latter.

3.4 Mean squared error

Previously, we motivated the ridge estimator as *i)* an ad hoc solution to collinearity, *ii)* a minimizer of a penalized sum of squares. An alternative motivation comes from studying the Mean Squared Error (MSE) of the ridge regression estimator. In general, for any estimator of a parameter $\boldsymbol{\theta}$:

$$\begin{aligned}
\text{MSE}(\hat{\boldsymbol{\theta}}) &= E[(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})^2] \\
&= \text{Var}(\hat{\boldsymbol{\theta}}) + [\text{Bias}(\hat{\boldsymbol{\theta}})]^2.
\end{aligned}$$

Hence, the MSE is a measure of the quality of the estimator. The MSE of the ridge estimator is:

$$\begin{aligned}
\text{MSE}[\hat{\boldsymbol{\beta}}(\lambda)] &= E[(\mathbf{W}_\lambda \hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} (\mathbf{W}_\lambda \hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\
&= E(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \mathbf{W}_\lambda \hat{\boldsymbol{\beta}}) - E(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_\lambda \hat{\boldsymbol{\beta}}) - E(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \boldsymbol{\beta}) + E(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}) \\
&= E(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \mathbf{W}_\lambda \hat{\boldsymbol{\beta}}) - E(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \mathbf{W}_\lambda \hat{\boldsymbol{\beta}}) - E(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \mathbf{W}_\lambda \boldsymbol{\beta}) + E(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \mathbf{W}_\lambda \boldsymbol{\beta}) \\
&\quad - E(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \mathbf{W}_\lambda \boldsymbol{\beta}) + E(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \mathbf{W}_\lambda \hat{\boldsymbol{\beta}}) + E(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \mathbf{W}_\lambda \boldsymbol{\beta}) \\
&\quad - E(\boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_\lambda \hat{\boldsymbol{\beta}}) - E(\hat{\boldsymbol{\beta}}^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \boldsymbol{\beta}) + E(\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}) \\
&= E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \mathbf{W}_\lambda (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})] \\
&\quad - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \mathbf{W}_\lambda \boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \mathbf{W}_\lambda \boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \mathbf{W}_\lambda \boldsymbol{\beta} \\
&\quad - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_\lambda \boldsymbol{\beta} - \boldsymbol{\beta}^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \boldsymbol{\beta} + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta} \\
&= E\big\{(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^{\mathrm{T}} \mathbf{W}_\lambda^{\mathrm{T}} \mathbf{W}_\lambda (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})\big\} + \boldsymbol{\beta}^{\mathrm{T}} (\mathbf{W}_\lambda - \mathbf{I})^{\mathrm{T}} (\mathbf{W}_\lambda - \mathbf{I}) \boldsymbol{\beta} \\
&= \sigma^2 \,\text{tr}\big\{\mathbf{W}_\lambda (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1} \mathbf{W}_\lambda^{\mathrm{T}}\big\} + \boldsymbol{\beta}^{\mathrm{T}} (\mathbf{W}_\lambda - \mathbf{I})^{\mathrm{T}} (\mathbf{W}_\lambda - \mathbf{I}) \boldsymbol{\beta}.
\end{aligned}$$

In the last step we have used $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2 [\mathbf{X}^T\mathbf{X}]^{-1})$ and the expectation of the quadratic form of a multivariate random variable $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\mu}_\varepsilon, \boldsymbol{\Sigma}_\varepsilon)$ is:

$$E(\boldsymbol{\varepsilon}^{\mathrm{T}} \boldsymbol{\Lambda} \boldsymbol{\varepsilon}) \quad = \quad \text{tr}(\boldsymbol{\Lambda} \boldsymbol{\Sigma}_\varepsilon) + \boldsymbol{\mu}_\varepsilon^{\mathrm{T}} \boldsymbol{\Lambda} \boldsymbol{\mu}_\varepsilon,$$

of course replacing $\boldsymbol{\varepsilon}$ by $\hat{\boldsymbol{\beta}}$ in this expectation. The first summand in the final derived expression for $\text{MSE}[\hat{\boldsymbol{\beta}}(\lambda)]$ is the sum of the variances of the ridge estimator, while the second summand can be thought of the "squared bias" of the ridge estimator.

Theorem 1. (DRAPER AND SMITH????)
There exists $\lambda > 0$ such that $\text{MSE}[\hat{\boldsymbol{\beta}}(\lambda)] < \text{MSE}[\hat{\boldsymbol{\beta}}(0)] = \text{MSE}[\hat{\boldsymbol{\beta}}]$.

Proof. To be included. Or refer to ???. □

The optimal choice of depends on unknown quantities $\boldsymbol{\beta}$ and σ^2 . In practice one applies cross-validation (see later). The data set is split many times into a training and test set. For each split the regression parameters are estimated for all choices of λ using the training data. Estimated parameters are evaluated on the test set. The λ that on average over the test sets performs best (in some sense) is selected.

Example 6. *Orthonormal design matrix*
Assume the design matrix $\mathbf{X}$ is orthonormal. Then, $\mathrm{MSE}[\hat{\boldsymbol{\beta}}] = p\,\sigma^2$ and

$$\mathrm{MSE}[\hat{\boldsymbol{\beta}}(\lambda)] \quad = \quad \frac{p\,\sigma^2}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2}\boldsymbol{\beta}^T\boldsymbol{\beta}.$$

The latter achieves its minimum at: $\lambda = p\sigma^2/\boldsymbol{\beta}^T\boldsymbol{\beta}$.

3.5 Bayesian regression

Ridge regression has a close connection to Bayesian linear regression. Bayesian linear regression assumes the parameters $\boldsymbol{\beta}$ and σ^2 to be the random variables. The conjugate priors for the parameters are:

$$\boldsymbol{\beta} \,|\, \sigma^2 \sim \mathcal{N}(\mathbf{0}, \frac{\sigma^2}{\lambda}\mathbf{I})$$

and

$$\sigma^2 \sim \mathcal{IG}(\alpha_0, \beta_0),$$

where $\mathcal{IG}$ denotes the inverse Gamma distribution with shape parameter α_0 and scale parameter β_0.

The posterior distribution of $\boldsymbol{\beta}$ and σ^2 is then:

$$\begin{aligned} & f_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{Y}, \mathbf{X}) \\ & \quad \propto \quad f_Y(\mathbf{Y} \,|\, \mathbf{X}, \boldsymbol{\beta}, \sigma^2) f_\beta(\boldsymbol{\beta}|\sigma^2) f_\sigma(\sigma^2) \\ & \quad \propto \quad \sigma^{-n} \exp\Big[-\frac{1}{2\sigma^2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^{\mathrm{T}}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})\Big] \\ & \qquad \times\ \sigma^{-p} \exp\Big[-\frac{\tau}{2\sigma^2}\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}\Big] \\ & \qquad \times\ [\sigma^2]^{-\alpha_0 - 1} \exp\Big[-\frac{\beta_0}{2\sigma^2}\Big]. \end{aligned}$$

This can be rewritten to:

$$f_{\boldsymbol{\beta},\sigma^2}(\boldsymbol{\beta}, \sigma^2 \,|\, \mathbf{Y}, \mathbf{X}) \quad \propto \quad g_{\boldsymbol{\beta}}(\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{Y}, \mathbf{X})\, g_{\sigma^2}(\sigma^2 \,|\, \mathbf{Y}, \mathbf{X})$$

where

$$\begin{aligned} & g_{\boldsymbol{\beta}}(\boldsymbol{\beta} \,|\, \sigma^2, \mathbf{Y}, \mathbf{X}) \ = \\ & \qquad \sigma^{-k} \exp\Big\{ -\frac{1}{2\sigma^2}\big[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\lambda)\big]^{\mathrm{T}}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})\big[\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}(\lambda)\big]\Big\}. \end{aligned}$$

Then, clearly the posterior mean of $\boldsymbol{\beta}$ is $E(\boldsymbol{\beta}) = \hat{\boldsymbol{\beta}}(\lambda)$. Hence, the ridge regression estimator can be viewed as a Bayesian estimate of $\boldsymbol{\beta}$ when imposing a Gaussian prior.

Show plot: The penalty parameter relates to the prior: a smaller penalty corresponds to wider prior, and a larger penalty to a more informative prior. Relate directly the variance in the prior of $\boldsymbol{\beta}$.

Include more details of derivation?

3.6 Constrained estimation

$$\min_{\boldsymbol{\beta}} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \quad = \quad \min_{\|\boldsymbol{\beta}\|_2^2 \leq \theta(\lambda)} \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2$$

Consider the following loss function:

$$\begin{aligned} \mathcal{L}(\boldsymbol{\beta};\lambda) &= \|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \\ &= \sum_{i=1}^{n}(Y_i - \mathbf{X}_{i*}\boldsymbol{\beta})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \end{aligned}$$

$$\begin{aligned} \frac{\partial}{\partial\boldsymbol{\beta}}\mathcal{L}(\boldsymbol{\beta};\lambda) &= -2\,\mathbf{X}^T(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + 2\,\lambda\,\mathbf{I}\,\boldsymbol{\beta} \\ &= -2\,\mathbf{X}^T\mathbf{Y} + 2\,(\mathbf{X}^T\mathbf{X} + \lambda\,\mathbf{I})\boldsymbol{\beta} \end{aligned}$$

3.7 Degrees of freedom

The degrees of freedom consumed by ridge regression is calculated. Recall from ordinary regression that:

$$\begin{aligned} \hat{\mathbf{Y}} &= \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \\ &= \mathbf{H}\mathbf{Y}, \end{aligned}$$

where $\mathbf{H}$ is called the hat matrix (as it put the 'hat' on $\mathbf{Y}$). The degrees of freedom used in the regression is then equal to $\mathrm{tr}(\mathbf{H})$, the trace of $\mathbf{H}$. In particular, if $\mathbf{X}$ is of full rank, i.e. $\mathrm{rank}(\mathbf{X}) = p$, then $\mathrm{tr}(\mathbf{H}) = p$.

By analogy, the ridge-version of the hat matrix is:

$$\mathbf{H}(\lambda) \quad = \quad \mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}$$

Continuing this analogy, the degrees of freedom of ridge regression is given by the trace of the hat matrix:

$$\begin{aligned} \mathrm{tr}[\mathbf{H}(\lambda)] &= \mathrm{tr}[\mathbf{X}(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^{\mathrm{T}}] \\ &= \sum_{j=1}^{p}\frac{d_{jj}^2}{d_{jj}^2 + \lambda} \end{aligned}$$

The degrees of freedom consumes by the regression is monotone decreasing in λ. In particular:

$$\lim_{\lambda\to\infty} \mathrm{tr}[\mathbf{H}(\lambda)] \quad = \quad 0.$$

That is, in the limit no information from $\mathbf{X}$ is used. Indeed, $\boldsymbol{\beta}$ is forced to equal $\mathbf{0}_{p\times 1}$ which is not derived from data.

3.8 Eigenvalue shrinkage

The effect of the ridge penalty may also studied from the perspective of singular values. Let the singular value decomposition of a matrix $\mathbf{X}$ be:

$$\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}},$$

where $\mathbf{D}$ an $n \times n$-diagonal matrix with the singular values, $\mathbf{U}$ an $n \times n$ dimensional matrix with columns containing the left singular vectors, and $\mathbf{V}$ a $p \times n$ dimensional matrix with columns containing the right singular vectors. **Point out orthogonality of U and V and their transposes. Plus notation of the vectors v_j . Mention dimension of X**

The OLS estimator can then be rewritten in terms of the SVD-matrices as:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \\
&= (\mathbf{V}\mathbf{D}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^{\mathrm{T}}\mathbf{Y} \\
&= (\mathbf{V}\mathbf{D}^2\mathbf{V}^{\mathrm{T}})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^{\mathrm{T}}\mathbf{Y} \\
&= \mathbf{V}\mathbf{D}^{-2}\mathbf{V}^{\mathrm{T}}\mathbf{V}\mathbf{D}\mathbf{U}^{\mathrm{T}}\mathbf{Y} \\
&= \mathbf{V}\mathbf{D}^{-2}\mathbf{D}\mathbf{U}^{\mathrm{T}}\mathbf{Y},
\end{aligned}$$

where $\mathbf{D}^{-2}\mathbf{D}$ is not simplified further to emphasize the effect of the ridge penalty. Similarly, the ridge estimator can be rewritten in terms of the SVD-matrices as:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_{p\times p})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \\
&= (\mathbf{V}\mathbf{D}\mathbf{U}^{\mathrm{T}}\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}} + \lambda\mathbf{I}_{p\times p})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^{\mathrm{T}}\mathbf{Y} \\
&= (\mathbf{V}\mathbf{D}^2\mathbf{V}^{\mathrm{T}} + \lambda\mathbf{V}\mathbf{V}^{\mathrm{T}})^{-1}\mathbf{V}\mathbf{D}\mathbf{U}^{\mathrm{T}}\mathbf{Y} \\
&= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I}_{n\times n})^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{V}\mathbf{D}\mathbf{U}^{\mathrm{T}}\mathbf{Y} \\
&= \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I}_{n\times n})^{-1}\mathbf{D}\mathbf{U}^{\mathrm{T}}\mathbf{Y}.
\end{aligned}$$

Combining the two results and writing $(\mathbf{D})_{jj} = d_{jj}$ we have:

$$d_{jj} \geq \frac{d_{jj}}{d_{jj}^2 + \lambda} \qquad \text{for all } \lambda > 0.$$

Thus, the ridge penalty shrinks the singular values.

Return to the problem of the super-collinearity of $\mathbf{X}$ in the high-dimensional setting ($p > n$). The super-collinearity implies the singularity of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ and prevents the calculation of the OLS estimator of the regression coefficients. However, $\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_{p\times p}$ is non-singular, with inverse:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_{p\times p})^{-1} = \sum_{j=1}^{p}(d_{jj}^2 + \lambda)^{-1}\mathbf{v}_j\mathbf{v}_j^{\mathrm{T}}.$$

The right-hand side is well-defined for $\lambda > 0$.

3.9 Efficient calculation

In the high-dimensional setting the number of covariates p is large compared to the number of samples n. In a microarray experiment $p = 40000$ and $n = 100$ is not uncommon. To perform ridge regression in this context, the expression needs to be evaluated numerically:

$$(\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda\mathbf{I}_{p\times p})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}.$$

For $p = 40000$ this requires the inversion of a 40000×40000 dimensional matrix. This is not feasible on most desktop computers. However, there is a workaround.

Revisit the singular value decomposition of $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$ and write $\mathbf{R} = \mathbf{U}\mathbf{D}$. As both $\mathbf{U}$ and $\mathbf{D}$ are $(n \times n)$-dimensional matrices, so is $\mathbf{R}$. Consequently, $\mathbf{X}$ is now decomposed as $\mathbf{X} = \mathbf{R}\mathbf{V}^{\mathrm{T}}$.

The ridge estimator can be rewritten in terms of $\mathbf{R}$ and $\mathbf{V}$:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= (\mathbf{X}^{\mathrm{T}}\mathbf{X} + \lambda \mathbf{I}_{p\times p})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y} \\
&= (\mathbf{V}\mathbf{R}^{\mathrm{T}}\mathbf{R}\mathbf{V}^{\mathrm{T}} + \lambda \mathbf{I}_{p\times p})^{-1}\mathbf{V}\mathbf{R}^{\mathrm{T}}\mathbf{Y} \\
&= (\mathbf{V}\mathbf{R}^{\mathrm{T}}\mathbf{R}\mathbf{V}^{\mathrm{T}} + \lambda \mathbf{V}\mathbf{V}^{\mathrm{T}})^{-1}\mathbf{V}\mathbf{R}^{\mathrm{T}}\mathbf{Y} \\
&= \mathbf{V}(\mathbf{R}^{\mathrm{T}}\mathbf{R} + \lambda \mathbf{I}_{n\times n})^{-1}\mathbf{V}^{\mathrm{T}}\mathbf{V}\mathbf{R}^{\mathrm{T}}\mathbf{Y} \\
&= \mathbf{V}(\mathbf{R}^{\mathrm{T}}\mathbf{R} + \lambda \mathbf{I}_{n\times n})^{-1}\mathbf{R}^{\mathrm{T}}\mathbf{Y}.
\end{aligned}$$

Hence, the reformulated ridge estimator involves the inversion of an $(n \times n)$-dimensional matrix. With $n = 100$ this is feasible on most standard computer.

Hastie and Tibshirani (2004) point out that the number of computation operations reduces from $\mathcal{O}(p^3)$ to $\mathcal{O}(pn^2)$. In addition, they point out that this computation short-cut can be used in combination with other loss functions, for instance that of a GLM **(refer to exercise)**. ∎

3.10 Simulation

3.11 Illustration

3.12 Exercises

3.13 Answers

4 Lasso regression

5 Graph theory

6 The multivariate normal distribution

A p-dimensional random variable $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_p)^{\mathrm{T}}$ following a multivariately normal distribution, denoted $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, has density:

$$f_{\mathbf{Y}}(\mathbf{y}) \quad = \quad (2\pi)^{-p/2} |\boldsymbol{\Sigma}|^{-1/2} \, \exp[-\frac{1}{2}(\mathbf{y} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})].$$

The density function of a multivariate normal random variable $\mathbf{Y}$ is often denoted by $\phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{y})$. The multivariate normal distribution has two parameters: the mean $\boldsymbol{\mu}$ and the covariance $\boldsymbol{\Sigma}$. The mean is defined as:

$$\begin{aligned} \boldsymbol{\mu} &= (\mu_1, \ldots, \mu_p)^{\mathrm{T}} \\ &= (E(Y_1), \ldots, E(Y_p))^{\mathrm{T}} \\ &= E(\mathbf{Y}) \\ &= \int_{\mathbb{R}^p} \mathbf{y} \, \phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{y}) \, d\mathbf{y}. \end{aligned}$$

The covariance matrix is a $p \times p$ positive definite matrix defined as:

$$\begin{aligned} \boldsymbol{\Sigma} &= \begin{pmatrix} \sigma_{11} & \sigma_{12} & \ldots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \ldots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \ldots & \sigma_{pp} \end{pmatrix} \\ &= \begin{pmatrix} \mathrm{Cov}(Y_1, Y_1) & \mathrm{Cov}(Y_1, Y_2) & \ldots & \mathrm{Cov}(Y_1, Y_p) \\ \mathrm{Cov}(Y_2, Y_1) & \mathrm{Cov}(Y_2, Y_2) & \ldots & \mathrm{Cov}(Y_2, Y_2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(Y_1, Y_1) & \mathrm{Cov}(Y_p, Y_2) & \ldots & \mathrm{Cov}(Y_p, Y_p) \end{pmatrix} \\ &= \mathrm{Cov}(\mathbf{Y}, \mathbf{Y}) \\ &= \int_{\mathbb{R}^p} (\mathbf{y} - \boldsymbol{\mu}) \, (\mathbf{y} - \boldsymbol{\mu})^{\mathrm{T}} \, \phi_{\boldsymbol{\mu},\boldsymbol{\Sigma}}(\mathbf{y}) \, d\mathbf{y}. \end{aligned}$$

Hence, the elements of $\boldsymbol{\Sigma}$ are given by:

$$\begin{aligned} (\boldsymbol{\Sigma})_{j_1,j_2} &= \sigma_{j_1,j_2} \\ &= \mathrm{Cov}(Y_{j_1}, Y_{j_2}) \\ &= E\{[Y_{j_1} - E(Y_{j_1})][Y_{j_2} - E(Y_{j_2})]\} \\ &= E\{[Y_{j_1} - \mu_{j_1}][Y_{j_2} - \mu_{j_2}]\} \\ &= E\{Y_{j_1} Y_{j_2} - \mu_{j_1} \mu_{j_2}]\}. \end{aligned}$$

Consequently, as $\sigma_{j_1,j_2} = \sigma_{j_2,j_1}$ the covariance matrix $\boldsymbol{\Sigma}$ is symmetric.

The covariance matrix is often parametrized differently as:

$$\begin{aligned}
\boldsymbol{\Sigma} &= \boldsymbol{\Upsilon}\mathbf{R}\boldsymbol{\Upsilon} \\
&:= \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p \end{pmatrix} \begin{pmatrix} 1 & \rho_{12} & \dots & \rho_{1p} \\ \rho_{21} & 1 & \dots & \rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \dots & 1 \end{pmatrix} \begin{pmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p \end{pmatrix} \\
&= \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} & \dots & \sigma_1\sigma_p\rho_{1p} \\ \sigma_2\sigma_1\rho_{21} & \sigma_2^2 & \dots & \sigma_2\sigma_p\rho_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_p\sigma_1\rho_{p1} & \sigma_p\sigma_2\rho_{p2} & \dots & \sigma_p^2 \end{pmatrix}.
\end{aligned}$$

The matrix $\mathbf{R}$ is called the correlation matrix for, as

$$\rho_{j_1,j_2} = \mathrm{Cov}(Y_{j_1}, Y_{j_2}) \,/\, \sqrt{\mathrm{Var}(Y_{j_1})\mathrm{Var}(Y_{j_1})},$$

we have:

$$\begin{aligned}
\mathbf{R} &= \mathrm{Corr}(\mathbf{Y}, \mathbf{Y})) \\
&= \begin{pmatrix} \mathrm{Corr}(Y_1, Y_1) & \mathrm{Corr}(Y_1, Y_2) & \dots & \mathrm{Corr}(Y_1, Y_p) \\ \mathrm{Corr}(Y_2, Y_1) & \mathrm{Corr}(Y_2, Y_2) & \dots & \mathrm{Corr}(Y_2, Y_2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Corr}(Y_1, Y_1) & \mathrm{Corr}(Y_p, Y_2) & \dots & \mathrm{Corr}(Y_p, Y_p) \end{pmatrix}.
\end{aligned}$$

Clearly, as $\sigma_j > 0$ for all $j = 1, \dots, p$, and thus $\boldsymbol{\Upsilon}$ is non-singular, one can always convert a covariance matrix to a correlation matrix by pre- and post-multiplication by $\boldsymbol{\Upsilon}^{-1}$:

$$\begin{aligned}
\mathbf{R} &= \boldsymbol{\Upsilon}^{-1}\boldsymbol{\Sigma}\boldsymbol{\Upsilon}^{-1} \\
&= \begin{pmatrix} \sigma_1^{-1} & 0 & \dots & 0 \\ 0 & \sigma_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p^{-1} \end{pmatrix} \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix} \begin{pmatrix} \sigma_1^{-1} & 0 & \dots & 0 \\ 0 & \sigma_2^{-1} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p^{-1} \end{pmatrix}.
\end{aligned}$$

One cannot recover $\boldsymbol{\Sigma}$ from $\mathbf{R}$ without knowledge of $\boldsymbol{\Upsilon}$.

Example 7.
A set of independent p random variables Y_j, distributed as $Y_j \sim \mathcal{N}(\mu_j, \sigma_j^2)$ for $j = 1, \dots, p$, can be written as a p-variate normal distribution:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_p \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \sigma_p^2 \end{pmatrix} \right).$$

The independence of the Y_j implies that $\mathrm{Cov}(Y_{j_1}, Y_{j_2}) = 0$, and thus that all off-diagonal element of $\boldsymbol{\Sigma}$ are zero.

Note
If $Y_1, \dots, Y_p$ are not independent, the vector $(Y_1, \dots, Y_p)^{\mathrm{T}}$ need not follow a multivariate normal distribution.

Example 8. *(The bivariate normal distribution)*
The special case $p = 2$ is called the bivariate normal distribution. A 2-dimensional random variable is bivariate normally distributed, denoted:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho_{12} \\ \sigma_1\sigma_2\rho_{12} & \sigma_2^2 \end{pmatrix}\right),$$

if its density is:

$$\begin{aligned} & f_{(Y_1,Y_2)}(y_1, y_2) \\ &= (2\pi)^{-1}[\sigma_1^2\sigma_2^2(1-\rho^2)]^{-1/2} \\ &\quad \times \exp\left\{-\frac{1}{2}[\sigma_1^2\sigma_2^2(1-\rho^2)]^{-1} \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} \sigma_2^2 & -\sigma_1\sigma_2\rho_{12} \\ -\sigma_1\sigma_2\rho_{12} & \sigma_1^2 \end{pmatrix} \begin{pmatrix} y_1-\mu_1 \\ y_2-\mu_2 \end{pmatrix}\right\}. \end{aligned}$$

The density of the bivariate normal distribution is visualized in Figure 1 for several choices of the covariance parameters.

6.1 The marginal and conditional distribution

The *marginal distribution* of a subset of the random variables $Y_1, \ldots, Y_p$ is the distribution of the random variables in the subset. More specifically, suppose the p variates can be divided into two exhaustive and mutually exclusive subsets $\mathcal{A}$ and $\mathcal{B}$, i.e.: *i)* $\mathcal{A}, \mathcal{B} \subset \{1, \ldots, p\}$, *ii)* $\mathcal{A} \cap \mathcal{B} = \emptyset$, *iii)* $\mathcal{A} \cup \mathcal{B} = \{1, \ldots, p\}$. Denote by $\mathbf{Y}_a$ and $\mathbf{Y}_b$ the random vectors that are obtained by restricting $\mathbf{Y}$ to the variates that correspond to the elements of subsets $\mathcal{A}$ and $\mathcal{B}$, respectively. The marginal density of $\mathbf{Y}_a$ is then:

$$f_{\mathbf{Y}_a}(\mathbf{y}_a) = \int_{\mathbb{R}^{|\mathcal{B}|}} f_{\mathbf{Y}_a,\mathbf{Y}_b}(\mathbf{y}_a, \mathbf{y}_b)\, d\mathbf{y}_b,$$

where $|\mathcal{B}|$ denotes the cardinality of the set $\mathcal{B}$.

Theorem 2.
If $\mathbf{Y} = (\mathbf{Y}_a^{\mathrm{T}}, \mathbf{Y}_b^{\mathrm{T}})^{\mathrm{T}}$ (with $\mathbf{Y}_a$ and $\mathbf{Y}_b$ as above) follows a multivariate normal distribution, i.e.:

$$\mathbf{Y} = \begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{bb} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}\right),$$

the marginal distribution of $\mathbf{Y}_a$ is $\mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$.

Proof. The proof is limited to $p = 2$ (for the general case we refer to **REFERENCE**). Let the 2-dimensional variable $(Y_1, Y_2)^{\mathrm{T}}$ follow a bivariate normal distribution. First normalize the variates individually: $\tilde{Y}_j = (Y_j - \mu_j)/\sigma_j$ for $j = 1, 2$. The joint density of $\tilde{Y}_1$ and $\tilde{Y}_2$ is then:

$$(2\pi)^{-1}[(1-\rho^2)]^{-1/2} \exp\left\{-\frac{1}{2}(1-\rho^2)^{-1} \begin{pmatrix} \tilde{Y}_1 \\ \tilde{Y}_1 \end{pmatrix}^{\mathrm{T}} \begin{pmatrix} 1 & -\rho \\ -\rho & 1 \end{pmatrix} \begin{pmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \end{pmatrix}\right\}.$$

Next apply the following change-of-variables:

$$\begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \begin{pmatrix} -C & -C \\ -C & C \end{pmatrix} \begin{pmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \end{pmatrix} \qquad \text{or} \qquad \begin{pmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \end{pmatrix} = -\frac{1}{2}C^{-1} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix},$$

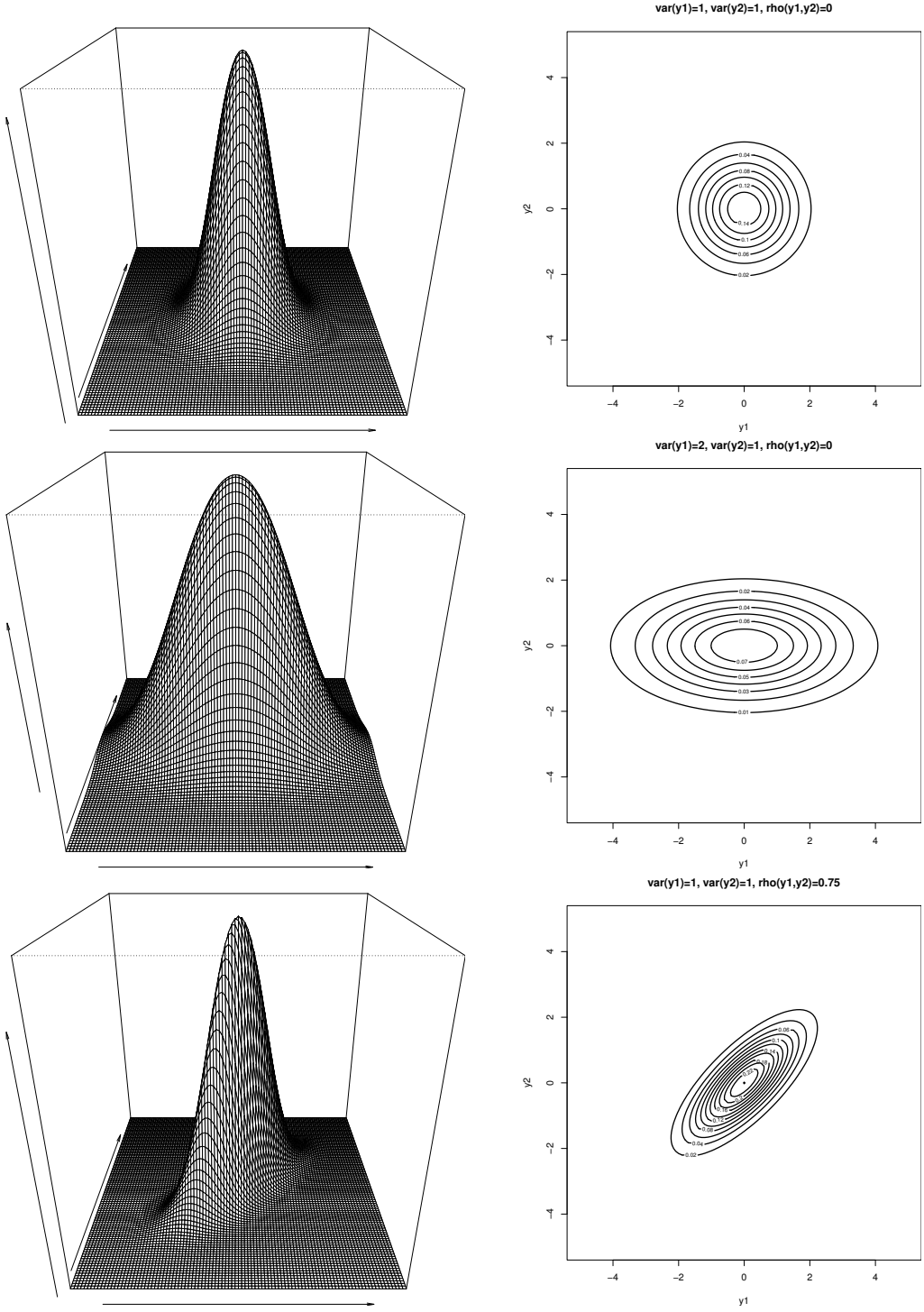

Figure 1: Three instances of the bivariate normal distribution. Left panels: 3d plots of the bivariate density; right panels: corresponding contour plots. Top row: $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho_{12} = 0$; middle row: $\sigma_1^2 = 2$, $\sigma_2^2 = 1$, $\rho_{12} = 0$; bottom row: $\sigma_1^2 = 1$, $\sigma_2^2 = 1$, $\rho_{12} = 0.75$.

with $C = \frac{1}{2}\sqrt{2}$. The Jacobian, the absolute value of determinant of the matrix with all first order partial derivatives, of this transformation (the one on the right) equals: $J = (2C)^{-2}$. The density function of (Z_1, Z_2) is:

$$\begin{aligned}
&\int_{-\infty}^{\infty} f_{(\tilde{Y}_1,\tilde{Y}_2)}(\tilde{y}_1,\tilde{y}_2)\, d\tilde{y}_1 \\
&\quad = \int_{-\infty}^{\infty} f_{(Z_1,Z_2)}\,(z_1,z_2)\, J\, dz_1 \\
&\quad = \int_{-\infty}^{\infty} (2\pi)^{-1}[(1-\rho^2)]^{-1/2}\,[2\,C]^{-2} \\
&\qquad\qquad \times \exp\left\{-[2C]^{-2}(1-\rho^2)^{-1}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}^{\mathrm{T}}\begin{pmatrix} 1-\rho & 0 \\ 0 & 1+\rho \end{pmatrix}\begin{pmatrix} z_1 \\ z_2 \end{pmatrix}\right\} dz_1 \\
&\quad = \int_{-\infty}^{\infty} (2\pi)^{-1}[(1-\rho^2)]^{-1/2}\,[2\,C]^{-2} \\
&\qquad\qquad \times \exp\left\{-[2C]^{-2}[(1-\rho)^{-1}z_1^2 + (1+\rho)^{-1}z_2^2]\right\}\, dz_1 \\
&\quad = \int_{-\infty}^{\infty} (2\pi)^{-1}[(1-\rho^2)]^{-1/2}\,[2\,C]^{-2} \\
&\qquad\qquad \times \exp\left\{-[2C]^{-2}[(1-\rho)^{-1}z_1^2]\right\}\, dz_1\, \exp\left\{-[2C]^{-2}(1+\rho)^{-1}z_2^2\right\}
\end{aligned}$$

The resulting bivariate random variable $\mathbf{Z}$ is now standard normally distributed: $\mathcal{N}(\mathbf{0}_{2\times 1}, \mathbf{I}_{2\times 2})$. The independence of Z_1 and Z_2 enables the factorization of the integral in the calculation of the marginal distribution. BLAH BLAH.

$$\begin{aligned}
f_{Y_1}(y_1) &= \int_{-\infty}^{\infty} f_{(Y_1,Y_2)}(y_1,y_2)\, dy_2 \\
&= (2\pi)^{-1/2}\sigma_1^{-1}\exp[-\frac{1}{2}(Y_1-\mu_1)^2/\sigma_1^2].
\end{aligned}$$

The result is now evident. □

The importance of Theorem lies in the fact that the marginal normal distribution is itself (multivariate) normal. Consequently, we can interpret the parameters of the multivariate normal in terms of the marginal variances and (bivariate) correlations. For instance, if $\mathbf{Y}$ follows a p-variate normal distribution, then:

$$\mu_1 \;=\; E(Y_1) \;=\; \int_{-\infty}^{\infty} y_1 f_{Y_1}(y_1)\, dy_1 \;=\; \int_{\mathbb{R}^p} y_1 f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y}.$$

Similar relations hold for the other parameters.

Theorem 3.
Suppose a p-variate normally distributed random variabele $\mathbf{Y}$ can be partitioned as follows:

$$\mathbf{Y} \;=\; \begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix} \;\sim\; \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}\right).$$

The conditional distribution of $\mathbf{Y}_a|\mathbf{Y}_b$ is then:

$$\mathbf{Y}|\mathbf{X} \;=\; \mathcal{N}(\boldsymbol{\mu}_Y + \boldsymbol{\Sigma}_{YZ}\boldsymbol{\Sigma}_{XX}^{-1}(\mathbf{X}-\boldsymbol{\mu}_X), \boldsymbol{\Sigma}_{YY} - \boldsymbol{\Sigma}_{YX}\boldsymbol{\Sigma}_{XX}^{-1}\boldsymbol{\Sigma}_{XY}).$$

Proof. Refer Theorem B.6.5 of Bickel and Doksum (2001) or **ANDERSON** □

Example 9. *Marginal vs. conditional distribution*
Consider the trivariate normal distribution:

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 2 & -1 & -1 \\ -1 & 3/2 & 1/2 \\ -1 & 1/2 & 3/2 \end{pmatrix} \right).$$

The marginal distribution of $(Y_2, Y_3)^{\mathrm{T}}$

$$\begin{pmatrix} Y_2 \\ Y_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 3/2 & 1/2 \\ 1/2 & 3/2 \end{pmatrix} \right).$$

To obtain the conditional distribution of $(Y_2, Y_3)^{\mathrm{T}}$ on Y_1 apply Theorem 3. For the conditional mean, we get:

$$\boldsymbol{\mu}_{(Y_2,Y_3)\,|\,Y_1} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} -1 \\ -1 \end{pmatrix} \left(\frac{1}{2}\right) (Y_1 - 0).$$

The conditional variance equals:

$$\boldsymbol{\Sigma}_{(Y_2,Y_3)\,|\,Y_1} = \begin{pmatrix} 3/2 & 1/2 \\ 1/2 & 3/2 \end{pmatrix} - \begin{pmatrix} -1 \\ -1 \end{pmatrix} \left(\frac{1}{2}\right) \begin{pmatrix} -1 \\ -1 \end{pmatrix}^{\mathrm{T}}.$$

The conditional distribution of $(Y_2, Y_3)^{\mathrm{T}}$ on Y_1 is thus:

$$\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \Big| \, Y_1 \sim \mathcal{N}\left(\begin{pmatrix} -\frac{1}{2}Y_1 \\ -\frac{1}{2}Y_1 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right).$$

Hence, conditional on Y_1, the random variable Y_2 and Y_3 are uncorrelated. The difference between the marginal and conditional distribution above is illustrated in Figure 2.

Biological interpretation.
Let the random variables Y_1, Y_2 and Y_3 represent expression levels of three genes. As Y_2 and Y_3 are independent given Y_1, their conditional independence graph is given by edge set $\mathcal{E} = \{Y_2 - Y_1, Y_1 - Y_3\}$. Would high expression levels of Y_2 be essential for the viability of the cell and Y_3 be a nuisance variable, one may neutralize the effect of Y_3 by controlling Y_1.

6.2 Estimation of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$

Consider an experiment involving a sample of n individuals of which a p traits are measured, denoted $\mathbf{y}_{*,i}$ for sample $= 1, \ldots, n$. Assuming the $\mathbf{y}_{*,i}$ are realizations from a multivariate normal distribution, the data may be used to estimate the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the distribution. To

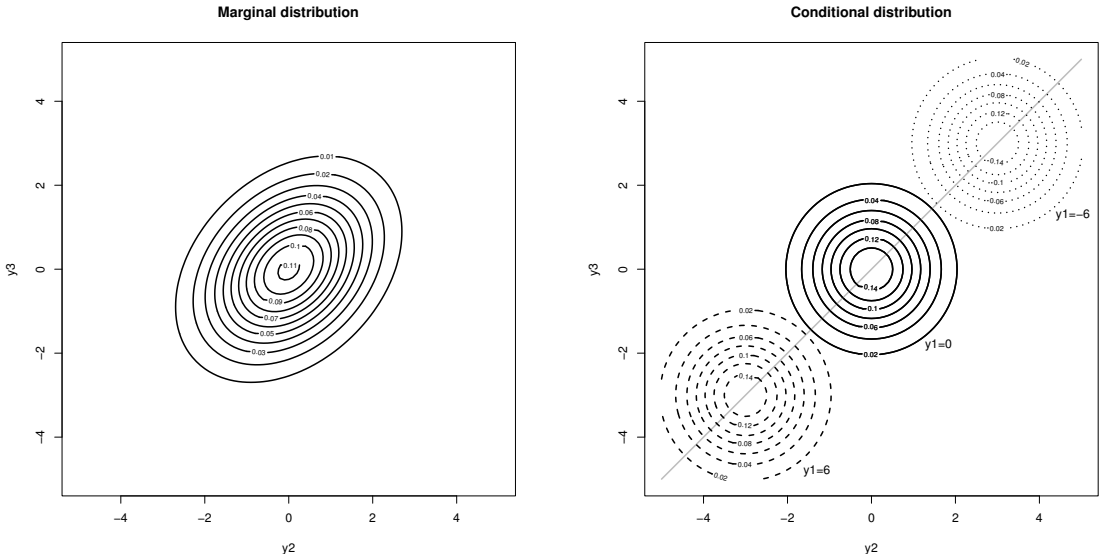

Figure 2: Marginal vs. conditional distribution. The marginal (left panel) and conditional (right panel) distribution of (Y_2, Y_3) (from Example 14) as contour plots.

derive the maximum likelihood estimate, we consider the log-likelihood:

$$\begin{aligned}
\mathcal{L}(\mathbf{Y}_{*,1}, \ldots, \mathbf{Y}_{*,n}) &= \sum_{i=1}^{n} \log[f_{\mathbf{Y}}(\mathbf{y}_{*,i})] \\
&\propto -\frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_{*,i}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}_{*,i}-\boldsymbol{\mu}) \\
&= -\frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}[(\mathbf{y}_{*,i}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}_{*,i}-\boldsymbol{\mu})] \\
&= -\frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\sum_{i=1}^{n}\mathrm{tr}[(\mathbf{y}_{*,i}-\boldsymbol{\mu})(\mathbf{y}_{*,i}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}] \\
&= -\frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{1}{2}\mathrm{tr}\Big[\sum_{i=1}^{n}(\mathbf{y}_{*,i}-\boldsymbol{\mu})(\mathbf{y}_{*,i}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\Big] \\
&= -\frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{n}{2}\mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}),
\end{aligned}$$

where

$$\mathbf{S} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_{*,i}-\boldsymbol{\mu})(\mathbf{y}_{*,i}-\boldsymbol{\mu})^{\mathrm{T}},$$

the sample covariance matrix. In the derivation of the log-likelihood we used *i)* $a = \mathrm{tr}(a)$ for a a scalar, *ii)* the cyclic property of the trace: $\mathrm{tr}(\mathbf{AB}) = \mathrm{tr}(\mathbf{BA})$, and *iii)* the additive property of the trace: $\mathrm{tr}(\mathbf{A}) + \mathrm{tr}(\mathbf{B}) = \mathrm{tr}(\mathbf{AB})$.

For the derivation of the ML estimate of $\boldsymbol{\mu}$ take the derivative with respect to $\boldsymbol{\mu}$:

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\mu}}\mathcal{L}(\mathbf{Y}_{*,1},\ldots,\mathbf{Y}_{*,n}) &\propto \sum_{i=1}^{n}\frac{\partial}{\partial \boldsymbol{\mu}}[(\mathbf{y}_{*,i}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{y}_{*,i}-\boldsymbol{\mu})] \\ &= -\sum_{i=1}^{n}[\boldsymbol{\Sigma}^{-1}(\mathbf{y}_{*,i}-\boldsymbol{\mu})]^{\mathrm{T}}+(\mathbf{y}_{*,i}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1} \\ &= -2\Big[\sum_{i=1}^{n}(\mathbf{y}_{*,i}-\boldsymbol{\mu})^{\mathrm{T}}\Big]\boldsymbol{\Sigma}^{-1},\end{aligned}$$

where we have used the chain rule and, for $\mathbf{A}$ independent of $\mathbf{y}$, that

$$\frac{\partial\,\mathbf{A}\mathbf{y}}{\partial\,\mathbf{y}} = \mathbf{A} \qquad \text{and} \qquad \frac{\partial\,\mathbf{y}^{\mathrm{T}}\mathbf{A}}{\partial\,\mathbf{y}} = \mathbf{A}^{\mathrm{T}}.$$

Equate the derivative to zero and solve for $\boldsymbol{\mu}$. This yields:

$$\hat{\boldsymbol{\mu}}_{ML} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{Y}_{*,i}.$$

This coincides with the univariate estimates of the mean.

The ML estimate of $\boldsymbol{\Sigma}$ is:

$$\arg\max_{\boldsymbol{\Sigma}>0}\frac{n}{2}\log(|\boldsymbol{\Sigma}^{-1}|)-\frac{n}{2}\mathrm{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1}).$$

To solve this apply the change-of variable $\boldsymbol{\Omega}=\boldsymbol{\Sigma}^{-1}$, which gives:

$$\arg\max_{\boldsymbol{\Omega}>0}\frac{n}{2}\log(|\boldsymbol{\Omega}|)-\frac{n}{2}\mathrm{tr}(\mathbf{S}\boldsymbol{\Omega}).$$

Equate the derivative with respect to $\boldsymbol{\Omega}$ to zero:

$$\frac{n}{2}\boldsymbol{\Omega}^{-1}-\frac{n}{2}\mathbf{S} = \mathbf{0}_{p\times p},$$

in which we have used that

$$\frac{\partial}{\partial\mathbf{X}}\log(|\mathbf{X}|) = \mathbf{X}^{-1} \qquad \text{and} \qquad \frac{\partial}{\partial\mathbf{X}}\mathrm{tr}(\mathbf{A}\mathbf{X}) = \mathbf{A}.$$

Solving this equation for $\boldsymbol{\Omega}$ yields the maximum likelihood estimate $\hat{\boldsymbol{\Sigma}}_{ML}=\mathbf{S}$.

Example 10. *Covariance estimation*
Estimation of the covariance matrix using the data from Table 1 yields:

$$\hat{\boldsymbol{\Sigma}} = \begin{matrix}\\ \texttt{MM1}\\ \texttt{MM2}\\ \texttt{TSG}\end{matrix}\begin{pmatrix} 0.2425 & 0.0812 & -0.0993 \\ 0.0812 & 0.5814 & -0.1866 \\ -0.0993 & -0.1866 & 0.2577 \end{pmatrix}.$$

From this estimate we may obtain the estimates of the correlation matrix:

$$\hat{\mathbf{R}} = \begin{pmatrix} 1.0000 & 0.2164 & -0.3972 \\ 0.2164 & 1.0000 & -0.4821 \\ -0.3972 & -0.4821 & 1.0000 \end{pmatrix}.$$

and of $\boldsymbol{\Upsilon}$: $\mathrm{diag}(\hat{\boldsymbol{\Upsilon}}) = c(0.4925, 0.7625, 0.5076)$. These estimates coincide with the univariate variance and bivariate correlation estimates, as suggested by Theorem 2.

6.3 Algebra with multivariate normally distributed random variables

Theorem 4.
Let $\mathbf{Y}$ be a p-variate normally distributed random variable: $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For any $q \times p$-dimensional matrix $\mathbf{A} \in \mathbb{R}^{q \times p}$, the random variable $\mathbf{AY}$ is distributed as $\mathcal{N}(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^{\mathrm{T}})$.

Proof. To be included. □

Theorem 5.
Let $\mathbf{Y}_a$ and $\mathbf{Y}_b$ be independent random variables distributed as $\mathbf{Y}_a \sim \mathcal{N}(\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})$ and $\mathbf{Y}_b \sim \mathcal{N}(\boldsymbol{\mu}_b, \boldsymbol{\Sigma}_{bb})$. Then:

$$\mathbf{Y}_a + \mathbf{Y}_b \quad \sim \quad \mathcal{N}(\boldsymbol{\mu}_a + \boldsymbol{\mu}_b, \boldsymbol{\Sigma}_{aa} + \boldsymbol{\Sigma}_{bb}).$$

Proof. To be included. □

6.4 Estimation of $\boldsymbol{\Sigma}$ when $p > n$

To be included: ad hoc (Schäfer and Strimmer, 2005, Ledoit and Wolf (2004), ridge (wij), lasso (Banerjee *et al.*, 2008, Friedman *et al.*, 2007).

6.5 Dataset

obs.	*MM1*	*MM2*	*TSG*	*obs.*	*MM1*	*MM2*	*TSG*
1	0.7069	3.0456	0.1329	35	0.9181	2.5301	-1.1091
2	1.2780	2.9740	-0.2307	36	1.0995	2.4074	-0.9819
3	1.0241	2.6215	-1.0268	37	0.2171	0.3928	0.5323
4	1.9692	2.7434	-0.9792	38	0.2134	1.7003	0.4777
5	1.7339	1.9604	-0.7639	39	1.3742	2.5333	-0.7047
6	1.1648	2.5200	0.0694	40	0.7518	2.6495	-0.9231
7	1.6664	2.0219	0.4182	41	0.3466	2.4796	-0.4910
8	0.7223	2.3734	-0.9818	42	0.4095	2.5608	-0.7833
9	1.3028	2.0927	-0.2982	43	0.8286	2.6012	0.2585
10	1.3662	1.8749	-0.4329	44	0.4735	1.8046	0.4952
11	1.3328	2.9184	-0.5327	45	1.2238	2.0156	-0.4167
12	0.6607	2.6653	-0.4462	46	-0.2194	1.3026	0.0120
13	2.2438	4.1512	-1.3046	47	1.2735	1.8488	-0.2404
14	1.6449	2.7535	-0.8217	48	1.5086	1.2532	-0.5658
15	1.2948	3.0529	-0.8477	49	1.0583	2.5911	-0.4010
16	1.4328	0.6720	-0.5411	50	1.0907	2.0880	0.2823
17	1.2793	4.0207	-0.4633	51	0.8655	1.7272	-1.0169
18	1.3623	2.2180	-0.2413	52	1.0067	1.6723	-0.4738
19	1.3164	3.8771	-0.4540	53	0.7097	3.0203	-0.6469
20	1.5871	3.5549	-1.1869	54	1.0356	2.3246	-0.5125
21	0.5472	3.5245	-0.8494	55	0.9834	2.0716	-0.8461
22	1.2153	4.4000	-0.7991	56	0.6392	2.4647	-0.9988
23	1.3725	2.4686	-0.8175	57	0.9944	2.3098	-0.6334
24	1.4675	2.0200	-0.4635	58	0.5905	2.8343	-0.2944
25	1.1766	2.8938	-0.8001	59	-0.0507	3.1692	-0.5449
26	2.0282	2.7562	-1.1482	60	1.1006	2.0447	-0.6471
27	1.6636	2.4145	-0.4318	61	0.7889	1.9683	0.0620
28	0.4312	3.7093	-1.0467	62	1.0756	1.8692	-0.3779
29	0.2680	3.4655	-0.4241	63	0.6907	1.9828	-0.2462
30	1.5317	2.7989	-1.2885	64	0.5442	1.0701	0.4336
31	1.4764	3.1314	-1.2885	65	0.6772	2.3257	0.1400
32	1.7428	3.3195	-1.6987	66	0.8593	1.9347	-0.1315
33	0.7137	1.7561	-0.2578	67	1.2610	2.4639	0.5910
34	1.0458	2.0934	-0.3535				

Table 1: Data set. Gene expression values of two methylation markers (MM1 and MM2) and a tumor suppressor gene (TSG). Data from a microarray experiment involving 67 samples.

7 Partial correlation

The partial correlation coefficient quantifies the correlation between two variables when conditioning on one or several other variables. More formally, let $\mathbf{Y} = (Y_1, \ldots, Y_p)^{\mathrm{T}}$ be a p-variate random variable, a, b elements of $\{1, \ldots, p\}$, $\mathcal{C}$ a subset of $\{1, \ldots, p\}$, Y_a and Y_b the corresponding variates of $\mathbf{Y}$, and $\mathbf{Y}_c$ obtained by limiting $\mathbf{Y}$ to the variates with an index in $\mathcal{C}$. The partial correlation coefficient between Y_a and Y_b conditional on $\mathbf{Y}_c$ is the correlation between the residuals of Y_a and Y_b after regressing them on $\mathbf{Y}_c$. Or,

$$\begin{aligned}
&\rho(Y_a, Y_b \,|\, \mathbf{Y}_c) \\
&\quad = \frac{\mathrm{Cov}(Y_a, Y_b \,|\, \mathbf{Y}_c)}{\sqrt{\mathrm{Var}(Y_a \,|\, \mathbf{Y}_c)}\,\sqrt{\mathrm{Var}(Y_b \,|\, \mathbf{Y}_c)}} \\
&\quad = \frac{E(Y_a, Y_b \,|\, \mathbf{Y}_c) - E(Y_a \,|\, \mathbf{Y}_c)\,E(Y_b \,|\, \mathbf{Y}_c)}{\sqrt{E(Y_a^2 \,|\, \mathbf{Y}_c) - E(Y_a \,|\, \mathbf{Y}_c)\,E(Y_a \,|\, \mathbf{Y}_c)}\,\sqrt{E(Y_b^2 \,|\, \mathbf{Y}_c) - E(Y_b \,|\, \mathbf{Y}_c)\,E(Y_b \,|\, \mathbf{Y}_c)}}. \qquad (2)
\end{aligned}$$

The order of the partial correlation coefficient is determined by the number of variables (that is, the cardinality of $\mathcal{C}$) it is conditional on.

A natural estimator of the partial correlation is obtained by replacing the conditional expectations in definition (2) by the corresponding estimates:

$$\hat{\rho}(Y_a, Y_b \,|\, \mathbf{Y}_c) \quad = \quad \frac{\sum e_{Y_a \,|\, \mathbf{Y}_c}\, e_{Y_b \,|\, \mathbf{Y}_c} - \big(\sum e_{Y_a \,|\, \mathbf{Y}_c}\big)\big(\sum e_{Y_b \,|\, \mathbf{Y}_c}\big)}{\sqrt{\sum e^2_{Y_a \,|\, \mathbf{Y}_c} - \big(\sum e_{Y_a \,|\, \mathbf{Y}_c}\big)^2}\,\sqrt{\sum e^2_{Y_b \,|\, \mathbf{Y}_c} - \big(\sum e_{Y_b \,|\, \mathbf{Y}_c}\big)^2}},$$

in which, e.g. $e_{Y_a \,|\, \mathbf{Y}_c} = Y_a - E(Y_a \,|\, \mathbf{Y}_c)$, the residual obtained when regressing Y_a on $\mathbf{Y}_c$.

Example 11. *(Illustration of partial correlation)*
Consider an artificial but not unrealistic situation in which two genes (referred to as gene 1 and gene 2) regulate a third (gene 3). Simulated expression levels of the three genes from five samples are given in Table 2.

obs.	*gene 1*	*gene 2*	*gene 3*	*res.* $2 \,|\, 1$	*res.* $3 \,|\, 1$
1	1.6067	1.2542	0.6955	0.1888	-0.0967
2	-2.1766	-1.7169	-0.1870	0.1282	-0.0724
3	0.7909	-0.6461	0.8966	-1.0839	0.2999
4	0.9350	0.2949	1.5312	-0.2538	0.9000
5	1.0485	1.6567	-0.3724	1.0207	-1.0308

Table 2: Artifical data set. Simulated gene expression values of three genes. The right two columns contain residuals from regressing the expression levels of gene 3 on gene 1 and gene 2 on gene 1.

The estimated correlation matrix, containing the partial correlations of order zero, for the three genes is:

$$\hat{\mathbf{R}} \quad = \quad \begin{pmatrix} 1.0000 & 0.8331 & 0.4542 \\ 0.8331 & 1.0000 & 0.0038 \\ 0.4542 & 0.0038 & 1.0000 \end{pmatrix}.$$

From the correlation matrix one concludes that gene 2 is marginally uncorrelated from gene 3. However, gene 1 and 2 are highly correlated (also marginally), indicating collinearity between the two genes. The collinearity may obscure the effect of gene 2 on gene 3. This can be assessed by the partial correlation. To calculate the partial correlation, regress the expression levels of gene 2 on gene 1 ($Y_{2,i} = \beta_{0,2} + \beta_{1,2}Y_{1,i} + \varepsilon_{2,i}$) and obtain the residuals $\hat{\varepsilon}_{2,i}$. Do the same for gene 3 (regress it on gene 1 and obtain the residuals). The residuals from both regression analyses are given in the two most right columns of Table 2. The correlation between the vectors of residuals equals -0.7604. This is the estimate of $\rho_{Y_2,Y_3\,|\,Y_1}$, the partial correlation of Y_2 and Y_3 conditioned on Y_1. Hence, although the expression levels of genes 2 and 3 are marginally uncorrelated, when controlling for gene 1 they are strongly correlated.

The partial correlation coefficient may be defined recursively. Let the zero-th order partial correlation coefficient be defined as:

$$\rho(Y_a, Y_b \,|\, \emptyset) \;=\; \frac{\mathrm{Cov}(Y_a, Y_b)}{\sqrt{\mathrm{Var}(Y_a)}\sqrt{\mathrm{Var}(Y_b)}}.$$

The higher order partial correlation coefficients are then defined recursively as:

$$\rho(Y_a, Y_b \,|\, \mathbf{Y}_c) \;=\; \frac{\rho(Y_a, Y_b \,|\, \mathbf{Y}_{c\setminus d}) - \rho(Y_a, Y_d \,|\, \mathbf{Y}_{c\setminus d})\,\rho(Y_d, Y_b \,|\, \mathbf{Y}_{c\setminus d})}{\sqrt{1-[\rho(Y_a, Y_d \,|\, \mathbf{Y}_{c\setminus d})]^2}\,\sqrt{1-[\rho(Y_d, Y_b \,|\, \mathbf{Y}_{c\setminus d})]^2}}, \tag{3}$$

where d is a element in $\mathcal{C}$.

Example 12. *(Estimation of the partial correlation, recursively)*
The partial correlation between genes 2 and 3 (conditional on gene 1) calculated recursively through formula (3) gives:

$$\begin{aligned}
\hat{\rho}(Y_2, Y_3 \,|\, Y_1) &= \frac{\hat{\rho}(Y_2, Y_3 \,|\, \emptyset) - \rho(Y_1, Y_2 \,|\, \emptyset)\,\hat{\rho}(Y_1, Y_3 \,|\, \emptyset)}{\sqrt{1-[\hat{\rho}(Y_1, Y_2 \,|\, \emptyset)]^2}\,\sqrt{1-[\hat{\rho}(Y_1, Y_3 \,|\, \emptyset)]^2}}, \\
&= \frac{0.0038 - 0.8331 \times 0.4542}{\sqrt{1-[0.8331]^2} \times \sqrt{1-[0.4542]^2}} \\
&= -0.7604.
\end{aligned}$$

This is indeed equal to the estimate obtained via the correlation between the residuals from regression analyses.

A third and final method for the calculation of the partial correlation coefficient is by inversion of the covariance matrix $\boldsymbol{\Sigma}$:

$$\rho(Y_a, Y_b \,|\, \mathbf{Y}_c) \;=\; \frac{-\left(\boldsymbol{\Sigma}^{-1}\right)_{a,b}}{\sqrt{\left(\boldsymbol{\Sigma}^{-1}\right)_{a,b}}\sqrt{\left(\boldsymbol{\Sigma}^{-1}\right)_{a,b}}} \;=\; \frac{-\left(\boldsymbol{\Omega}\right)_{a,b}}{\sqrt{\left(\boldsymbol{\Omega}\right)_{a,b}}\sqrt{\left(\boldsymbol{\Omega}\right)_{a,b}}}. \tag{4}$$

Before we prove this equality, we first illustrate it on Example **??**.

Example 13. *(Estimation of the partial correlation via the covariance matrix inversion)*
The estimate of the precision matrix $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is:

$$\hat{\boldsymbol{\Omega}} \;=\; \begin{pmatrix} 4.3665 & -3.9317 & -3.7327 \\ -3.9317 & 4.0646 & 3.3576 \\ -3.7327 & 3.3576 & 4.7967 \end{pmatrix}.$$

Rescale $\mathbf{\Omega}$ to have a unit diagonal and multiply by -1 yields the partial correlation matrix:

$$\hat{\mathbf{\Omega}} \quad = \quad \begin{pmatrix} 1.0000 & 0.9333 & 0.8156 \\ 0.9333 & 1.0000 & -0.7604 \\ 0.8156 & -0.7604 & 1.0000 \end{pmatrix}.$$

Also this estimate of the $\rho(Y_2, Y_3 \,|\, Y_1)$ equals -0.7604.

7.1 Inverse variance lemma

The key result behind the relation between the partial correlation coefficient and the precision matrix, as specified by Formula (4), is the Inverse Variance Lemma. This lemma specifies the precision matrix of partitioned multivariate random variable.

Lemma 1. (Inverse Variance Lemma)
Let a p-variate normal random variable $\mathbf{Y}$ be partitioned as $\mathbf{Y} = (\mathbf{Y}_a^{\mathrm{T}} \mathbf{Y}_b^{\mathrm{T}})^{\mathrm{T}}$ where $\mathbf{Y}_a$ and $\mathbf{Y}_b$ are p_a- and p_b-dimensional such that $p_a + p_b = p$. The inverse of the partitioned variance $\mathrm{Var}(\mathbf{Y}) = \mathrm{Var}[(\mathbf{Y}_a^{\mathrm{T}}, \mathbf{Y}_b^{\mathrm{T}})^{\mathrm{T}}]$ is:

$$\left\{ \mathrm{Var}\left[\begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix} \right] \right\}^{-1} \quad = \quad \begin{pmatrix} [\mathrm{Var}(\mathbf{Y}_a)]^{-1} + \mathbf{B}^{\mathrm{T}} [\mathrm{Var}(\mathbf{Y}_b \,|\, \mathbf{Y}_a)]^{-1} \mathbf{B} & -\mathbf{B}^{\mathrm{T}} [\mathrm{Var}(\mathbf{Y}_b \,|\, \mathbf{Y}_a)]^{-1} \\ -[\mathrm{Var}(\mathbf{Y}_b \,|\, \mathbf{Y}_a)]^{-1} \mathbf{B} & [\mathrm{Var}(\mathbf{Y}_b \,|\, \mathbf{Y}_a)]^{-1} \end{pmatrix},$$

where $\mathbf{B} = \mathrm{Cov}(\mathbf{Y}_a, \mathbf{Y}_b)[\mathrm{Var}(\mathbf{Y}_a)]^{-1}$.

Proof. By assumption:

$$\mathbf{Y} \quad = \quad \begin{pmatrix} \mathbf{Y}_a \\ \mathbf{Y}_b \end{pmatrix} \quad \sim \quad N\left(\begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}, \begin{pmatrix} \mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb} \end{pmatrix} \right).$$

Application of Theorem 6, which provides an explicit expression for the inverse of a 2×2 symmetric block matrix, gives:

$$\begin{pmatrix} \mathbf{\Sigma}_{aa} & \mathbf{\Sigma}_{ab} \\ \mathbf{\Sigma}_{ba} & \mathbf{\Sigma}_{bb} \end{pmatrix} \quad = \quad \begin{pmatrix} \mathbf{\Sigma}_{aa}^{-1} + \mathbf{F}\, \mathbf{E}^{-1}\, \mathbf{F}^T & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\, \mathbf{F}^T & \mathbf{E}^{-1}, \end{pmatrix},$$

where $\mathbf{E} = \mathbf{\Sigma}_{bb} - \mathbf{\Sigma}_{ba}\, \mathbf{\Sigma}_{aa}^{-1}\, \mathbf{\Sigma}_{ab}$ and $\mathbf{F} = \mathbf{\Sigma}_{aa}^{-1}\, \mathbf{\Sigma}_{ba}$. From Theorem 3, which specifies the conditional distribution of $\mathbf{Y}_b$ on $\mathbf{Y}_a$, it follows that $\mathbf{E} = \mathrm{Var}(\mathbf{Y}_b \,|\, \mathbf{Y}_a)$. After noting that $\mathbf{\Sigma}_{aa} = \mathrm{Var}(\mathbf{Y}_a)$, $\mathbf{\Sigma}_{bb} = \mathrm{Var}(\mathbf{Y}_b)$ and $\mathbf{\Sigma}_{ab} = \mathrm{Cov}(\mathbf{Y}_a, \mathbf{Y}_b)$, the lemma is evident. □

The main take-away of the Inverse Variance Lemma is to be found in the term in the bottom right corner of the precision matrix: (the inverse of) the covariance matrix of $\mathbf{Y}_b$ conditional on $\mathbf{Y}_a$ can be obtained via inversion of the joint covariance matrix of $\mathbf{Y}_a$ and $\mathbf{Y}_b$. The following two corollaries make this explicit.

Corollary 1. (Corollary 5.8.1, Whittaker (1990))
Each diagonal element of the inverse covariance matrix is the reciprocal of a partial variance.

Proof. In the Inverse Variance Lemma set $p_2 = 1$. Then:

$$[\mathrm{Var}(Y_b \,|\, \mathbf{Y}_a)]^{-1} \quad = \quad 1/\mathrm{Var}(Y_b \,|\, \mathbf{Y}_a)$$

for this is a scalar. This holds for any diagonal element as any can be selected by permutation of the original random vector. □

Corollary 2. (Corollary 5.8.2, Whittaker (1990))
Each off-diagonal element of the inverse variance matrix (scaled to have a unit diagonal) is the negative of the partial correlation between the two corresponding variables, conditioned on all remaining variables.

Proof. In the Inverse Variance Lemma set $p_2 = 2$ and write $\mathbf{Y}_b = (Y_{b_1}, Y_{b_2})^{\mathrm{T}}$. Then:

$$[\mathrm{Var}(\mathbf{Y}_b \,|\, \mathbf{Y}_a)]^{-1} \;=\; (\mathbf{\Sigma}_{b|a})^{-1} \;=\; \mathbf{\Omega}_{b|a} \;:=\; \begin{pmatrix} \omega_{b_1,b_1} & \omega_{b_1,b_2} \\ \omega_{b_2,b_1} & \omega_{b_2,b_2} \end{pmatrix}.$$

Turning this around yields:

$$\begin{aligned} \begin{pmatrix} \mathrm{Var}(Y_{b_1} \,|\, \mathbf{Y}_a) & \mathrm{Cov}(Y_{b_1}, Y_{b_2} \,|\, \mathbf{Y}_a) \\ \mathrm{Cov}(Y_{b_1}, Y_{b_2} \,|\, \mathbf{Y}_a) & \mathrm{Var}(Y_{b_2} \,|\, \mathbf{Y}_a) \end{pmatrix} &= \begin{pmatrix} \omega_{b_1,b_1} & \omega_{b_1,b_2} \\ \omega_{b_2,b_1} & \omega_{b_2,b_2} \end{pmatrix}^{-1} \\ &= \frac{1}{\omega_{b_1,b_1}\omega_{b_2,b_2} - \omega_{b_1,b_2}\omega_{b_2,b_1}} \begin{pmatrix} \omega_{b_2,b_2} & -\omega_{b_2,b_1} \\ -\omega_{b_1,b_2} & \omega_{b_1,b_1} \end{pmatrix}. \end{aligned}$$

Rescale this to have a unit diagonal and one obtains:

$$\frac{-\omega_{b_1,b_2}}{\sqrt{\omega_{b_1,b_1}\omega_{b_2,b_2}}} \;=\; \frac{\mathrm{Cov}(Y_{b_1}, Y_{b_2} \,|\, \mathbf{Y}_a)}{\sqrt{\mathrm{Var}(Y_{b_1} \,|\, \mathbf{Y}_a)}\,\sqrt{\mathrm{Var}(Y_{b_2} \,|\, \mathbf{Y}_a)}},$$

which is the partial correlation coefficient. Permutation extends the result to all off-diagonal elements. □

It is this corollary that underpins Formula (4). The corollary is illustrated using the follow analytic example.

Example 14. *Calculation of the partial correlation via the inverse covariance matrix.*
Consider the 3×3 covariance matrix (without loss of generality rescaled to be a correlation matrix):

$$\mathbf{\Sigma} \;=\; \begin{pmatrix} 1 & \rho_{12} & \rho_{13} \\ \rho_{12} & 1 & \rho_{23} \\ \rho_{13} & \rho_{23} & 1 \end{pmatrix}.$$

The corresponding precision matrix is then:

$$\mathbf{\Sigma}^{-1} \;=\; \frac{1}{\det(\mathbf{\Sigma})} \begin{pmatrix} 1-\rho_{23}^2 & \rho_{13}\,\rho_{23} - \rho_{12} & \rho_{12}\,\rho_{23} - \rho_{13} \\ \rho_{13}\,\rho_{23} - \rho_{12} & 1-\rho_{13}^2 & \rho_{12}\,\rho_{13} - \rho_{23} \\ \rho_{12}\,\rho_{23} - \rho_{13} & \rho_{12}\,\rho_{13} - \rho_{23} & 1-\rho_{12}^2 \end{pmatrix}.$$

Let $a = \{3\}$, $b = \{1, 2\}$, $Y_a = Y_3$ and $\mathbf{Y}_b = (Y_1, Y_2)^{\mathrm{T}}$. Then, up a factor of the determinant of $\mathbf{\Sigma}$, the right submatrix of the precision matrix is identical to:

$$\mathrm{Var}(\mathbf{Y}_b \,|\, Y_a) \;=\; \mathbf{\Sigma}_{bb} - \mathbf{\Sigma}_{ba}\,\mathbf{\Sigma}_{aa}^{-1}\,\mathbf{\Sigma}_{ab} \;=\; \begin{pmatrix} 1 & \rho_{12} \\ \rho_{12} & 1 \end{pmatrix} - \begin{pmatrix} \rho_{13}^2 & \rho_{13}\,\rho_{23} \\ \rho_{13}\,\rho_{23} & \rho_{23}^2. \end{pmatrix},$$

where expression of the conditional variance stems from Theorem 3. The factor cancels out when calculating the partial correlation coefficient:

$$\rho(Y_1, Y_2 \,|\, Y_3) \;=\; \frac{\left(\mathbf{\Sigma}^{-1}\right)_{12}}{\sqrt{\left(\mathbf{\Sigma}^{-1}\right)_{11}}\,\sqrt{\left(\mathbf{\Sigma}^{-1}\right)_{22}}} \;=\; \frac{\rho_{13}\,\rho_{23} - \rho_{12}}{\sqrt{1-\rho_{23}^2}\,\sqrt{1-\rho_{13}^2}}.$$

This is exactly the result implied by Corollary 2.

7.2 Partial correlation and conditional independence

The relevance of the partial correlation coefficient with respect to the reconstruction of conditional independence graphs is expressed by the next corollary.

Corollary 3. (From Whittaker)
Let $\mathbf{Y}_a$, $\mathbf{Y}_b$ and $\mathbf{Y}_c$ by p_a-, p_b- and p_c-dimensional normally distributed random vectors. Then, $\mathbf{Y}_a$ and $\mathbf{Y}_b$ are independent conditional on $\mathbf{Y}_c$, i.e. $\mathbf{Y}_a \perp \mathbf{Y}_b \,|\, \mathbf{Y}_c$, if and only if:

$$(\mathbf{\Omega})_{a,b} \quad = \quad \mathbf{0}_{|\mathcal{A}|\times|\mathcal{B}|},$$

where $\mathbf{\Omega} = \{\mathrm{Var}[(\mathbf{Y}_a, \mathbf{Y}_b, \mathbf{Y}_c)]\}^{-1}$.

Corollary 3 gives a simple parametric criterion for (conditional) pairwise independence: a zero partial covariance, or (after rescaling) a zero partial correlation, corresponds to the absence of the corresponding edge in the conditional independence graph.

Example 15. *Equivalence of a zero partial correlation and conditional independence.*
Consider the 3×3 covariance matrix (without loss of generality rescaled to be a correlation matrix) of a trivariate normal distribution (as in Example 14). Its inverse $\mathbf{\Omega}$, the precision matrix, is then as in Example 14. Now assume $(\mathbf{\Omega})_{13} = \rho_{12}\rho_{23} - \rho_{13} = 0$. Corollary 3 then says Y_1 and Y_3 are independent conditional on Y_2: $Y_1 \perp Y_3 \,|\, Y_2$. This suggests that the joint density of Y_1, Y_2 and Y_3 would factorize as implied by Proposition **WELKE DAN?**. This is indeed the case, confer:

$$\begin{aligned}
& f_{(Y_1,Y_2,Y_3)}(y_1,y_2,y_3) \\
&\quad = \quad C\,\exp(-\frac{1}{2}\mathbf{y}^{\mathrm{T}}\mathbf{\Omega}\mathbf{y}) \\
&\quad = \quad C\,\exp\{-\frac{1}{2}[(\mathbf{\Omega})_{11}\,y_1^2 + (\mathbf{\Omega})_{22}\,y_2^2 + (\mathbf{\Omega})_{33}\,y_3^2 + 2\,(\mathbf{\Omega})_{12}\,y_1y_2 + 2\,(\mathbf{\Omega})_{23}\,y_2y_3]\} \\
&\quad = \quad C\,\exp\{-\frac{1}{2}[(\mathbf{\Omega})_{11}\,y_1^2 + 2\,(\mathbf{\Omega})_{12}\,y_1y_2 + (\mathbf{\Omega})_{22}\,y_2^2]\} \\
&\qquad\quad \times \exp\{-\frac{1}{2}[2\,(\mathbf{\Omega})_{33}\,y_3^2 + (\mathbf{\Omega})_{23}\,y_2y_3]\} \\
&\quad = \quad g(y_1,y_2)\,h(y_3,y_2),
\end{aligned}$$

for some suitably chosen functions $g(\cdot,\cdot)$ and $h(\cdot,\cdot)$.

7.3 Relation to regression

8 A single molecular level

8.1 Experimental design, data, and notation

We consider integrative oncogenomics studies with a time-course set-up that aim to elucidate the molecular mechanisms within the cancer cell. To this end an integrative oncogenomics study profiles *multiple* molecular levels of the same sample in high-throughput fashion using microarrays. The time-course set-up implies that each sample included in the study is followed over time and, at multiple time points during this period, interrogated molecularly.

In the experiment we assume that n samples are followed over time. The expression levels of p genes of sample i are measured at $\mathcal{T}$ time points. The time points are assumed to be identical for each sample. Let $Y_{j,i,t}$ be the zero-centered expression level of gene j in sample i at time point t. The matrix of all gene expression profiles of all samples at time point t is denoted:

$$\mathbf{Y}_{*,*,t} = \left(\mathbf{Y}_{*,1,t} \,|\, \mathbf{Y}_{*,2,t} \,|\, \ldots \,|\, \mathbf{Y}_{*,n,t}\right) = \begin{pmatrix} Y_{1,1,t} & Y_{1,2,t} & \ldots & Y_{1,n,t} \\ Y_{2,1,t} & Y_{2,2,t} & \ldots & Y_{2,n,t} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{p,1,t} & Y_{p,2,t} & \ldots & Y_{p,n,t} \end{pmatrix}$$

Similarly, the matrix of all gene expression profiles of sample i is written as:

$$\mathbf{Y}_{*,i,*} = \left(\mathbf{Y}_{*,i,1} \,|\, \mathbf{Y}_{*,i,2} \,|\, \ldots \,|\, \mathbf{Y}_{*,i,\mathcal{T}}\right) = \begin{pmatrix} Y_{1,i,1} & Y_{1,i,2} & \ldots & Y_{1,i,\mathcal{T}} \\ Y_{2,i,1} & Y_{2,i,2} & \ldots & Y_{2,i,\mathcal{T}} \\ \vdots & \vdots & \ddots & \vdots \\ Y_{p,i,1} & Y_{p,i,2} & \ldots & Y_{p,i,\mathcal{T}} \end{pmatrix}$$

At each time point each sample is also interrogated genomically. This yields a collection of DNA copy number profiles. An element of these profiles is denoted X_{jit}, the DNA copy number corresponding to gene j in sample i and time point t. In analogous fashion as above we write the matrix of the samples' DNA copy number data at a particular time point, and the matrix of genomic profiles of a particular sample over time.

Alternatively, $X_{j,i,t}$ may represent microRNA expression. Notice that the matrix $\mathbf{X}_{*,*,t}$ may then be of different dimensions, that is, may have a different number of rows.

The typical sample size of time-course experiments in terms of the number of time points sampled hardly ever exceeds ten. Usually, the number of cell lines involved in the experiments is also very small, say, a handful maximum. **(ref for these numbers.)**

8.2 The VAR(1) model

The gene expression data from the time-course experiment is modeled by a VAR(1) (first-order vector autoregressive) process:

$$\mathbf{Y}_{*,i,t} \,|\, \mathbf{Y}_{*,i,t-1}, \ldots, \mathbf{Y}_{*,i,1} = \boldsymbol{\nu} + \mathbf{A}\mathbf{Y}_{*,i,t-1} + \boldsymbol{\varepsilon}_{*,i,t}, \tag{5}$$

where $\boldsymbol{\nu}$ the $p \times 1$ intercept vector, $\mathbf{A}$ a $p \times p$ coefficient matrix, and $\boldsymbol{\varepsilon}_{*,i,t}$ a $p \times 1$ vector with the errors. Throughout it is assumed that $\boldsymbol{\varepsilon}_{*,i,t} \sim \mathcal{N}(\mathbf{0}_{p\times 1}, \boldsymbol{\Sigma}_{\varepsilon})$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}_{*,i_1,t_1}, \boldsymbol{\varepsilon}_{*,i_2,t_2}) = \mathbf{0}$ if $t_1 \neq t_2$ or $i_1 \neq i_2$. Thus, for a 3-gene pathway model (5) becomes:

$$\begin{cases} Y_{1,i,t} = \nu_1 + a_{11}Y_{1,i,t-1} + a_{12}Y_{2,i,t-1} + a_{13}Y_{3,i,t-1} + \varepsilon_{1,i,t}, \\ Y_{2,i,t} = \nu_2 + a_{21}Y_{1,i,t-1} + a_{22}Y_{2,i,t-1} + a_{23}Y_{3,i,t-1} + \varepsilon_{2,i,t}, \\ Y_{3,i,t} = \nu_3 + a_{31}Y_{1,i,t-1} + a_{32}Y_{2,i,t-1} + a_{33}Y_{3,i,t-1} + \varepsilon_{3,i,t}, \end{cases}$$

where for notational simplicity we have dropped the conditioning on the past at the left hand-side of the equations.

VAR(1) models have already been applied to modeling gene expression data from a time-course experiment in Fujita *et al.* (2007). This comprises a straightforward application of a VAR(1) model with a lasso penalty.

Also Abegaz and Wit (2013) applied it. Summarize briefly.

Example 16. *Illustration of the VAR(1) model.* Consider the feed-forward motif, apparently a common motif in the transcriptional network (Alon, 2007). The feed-forward motif comprises three genes. Let the random variables Y_1, Y_2 and Y_3 represent their gene expression levels. The feed-forward motif is then defined by the edge set $\mathcal{V} = \{Y_1 \to Y_2, Y_1 \to Y_3, Y_2 \to Y_3\}$. Over both paths from Y_1 to Y_3 the overall sign should be the same. The motif is depicted in Figure 3.

The particular version of the feed-forward motif used to generate the data depicted in Figure 4 is given by the following model:

$$\begin{cases} Y_{1,i,t} &= & & & & & \varepsilon_{1,i,t}, \\ Y_{2,i,t} &= & -\frac{5}{2}Y_{1,i,t-1} & & & + & \varepsilon_{2,i,t}, \\ Y_{3,i,t} &= & \frac{9}{5}Y_{1,i,t-1} & - & \frac{3}{2}Y_{2,i,t-1} & + & \varepsilon_{3,i,t}. \end{cases} \tag{6}$$

in which $\mathbf{\Sigma}_\varepsilon = \frac{1}{4}\mathbf{I}_{p\times p}$.

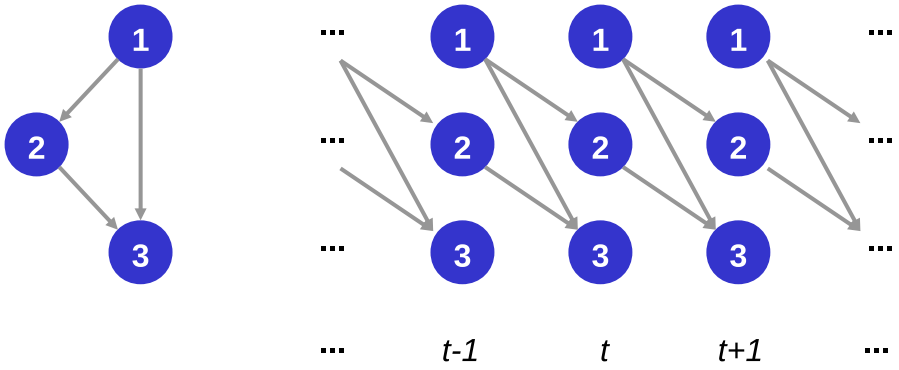

Figure 3: On the left-hand side: schemata of the feed-forward motif. To the right, the feed-forward motif unfolded.

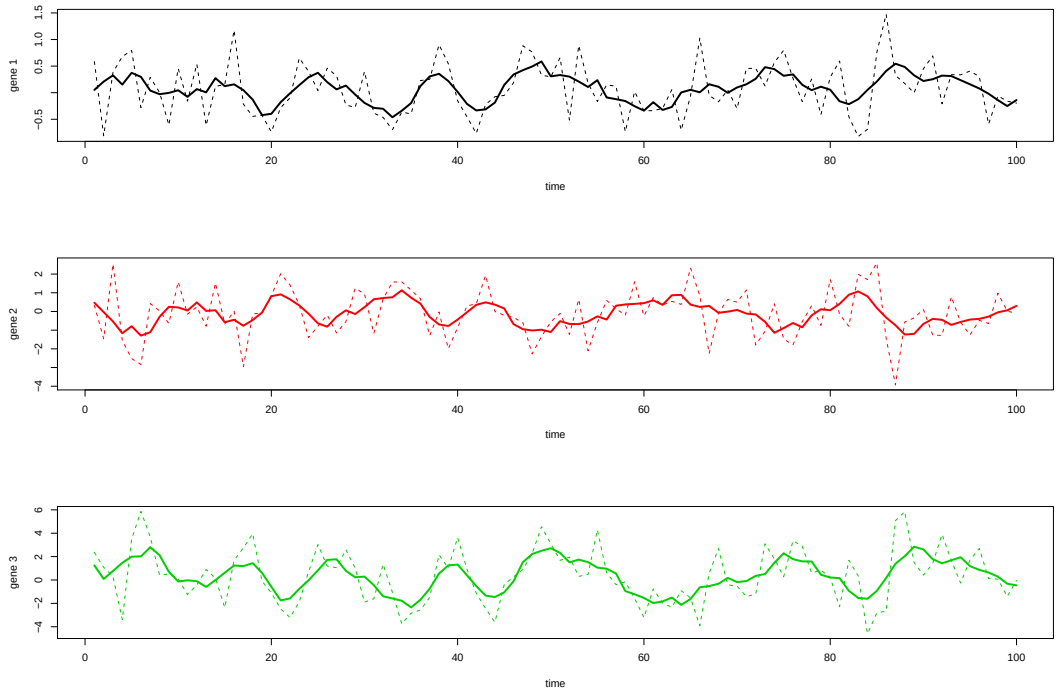

Figure 4: Data from the feed-forward example as specified by model (6). The dashed lines connect the data points, the solid line is a smoothing of the data to emphasize the trend.

8.2.1 Stability and stationarity

The VAR(1) process is said to be *stable* if all eigenvalues of $\mathbf{A}$ have modulus less than one, i.e. $|\lambda_j(\mathbf{A})| < 1$ for $j = 1, \ldots, p$. This is equivalent to:

$$\det(\mathbf{I}_{p\times p} - \mathbf{A}\,\mathbf{x}) \neq 0 \qquad \text{for } |\mathbf{x}| \leq 1.$$

The process may also be well-defined if the stability does not hold.

The VAR(1) process is said to be *strictly stationary* if $P(\mathbf{Y}_{*,i,1}, \ldots, \mathbf{Y}_{*,i,t}) = P(\mathbf{Y}_{*,i,\tau}, \ldots, \mathbf{Y}_{*,i,t+\tau})$ for any $\tau \in \mathbb{N}$. In words, the joint distribution (i.e. $\boldsymbol{\mu}_y$ and $\boldsymbol{\Sigma}_y$) does not change over time. In case the definition of strictly stationary is relaxed: the first two moments do not dependent on t, the process is referred to as *stationary*.

The two concepts are related by Proposition 2.1 of Lütkepohl (2005): model (5) yields a stationary time series if it is stable.

Example 17. In Example 16 the eigenvalues of $\mathbf{A}$ are $\lambda_j(\mathbf{A}) = 0$ for $j = 1, \ldots, p$. Hence, the process is stable, and therefore stationary.

8.2.2 The MA representation

Assume $\mathbf{Y}_{*,i,t}$ follows a VAR(1) model. Then, assuming the model is stable, the process can be represented by:

$$\begin{aligned}\mathbf{Y}_{*,i,t} &= \boldsymbol{\mu} + \sum_{t_0=0}^{\infty} \mathbf{A}^{t_0}\boldsymbol{\varepsilon}_{t-t_0} \\ &= \boldsymbol{\mu} + \sum_{t_0=0}^{\infty} \boldsymbol{\Phi}_{t_0}\boldsymbol{\varepsilon}_{t-t_0}.\end{aligned}$$

This is called the *moving average (MA) representation* (or sometimes *canonical*, *fundamental* or *prediction error representation*). The representation expresses $\mathbf{Y}_{*,i,t}$ in past and present error (innovations). The $\boldsymbol{\Phi}_{t_0}$ are referred to as the moving average coefficient matrices. The key motivation for this representation is the (more explicit) relation to the mean and autocovariance of the process.

8.2.3 The structural VAR model

The *structural VAR(1) model* is a reformulation of the canonical VAR(1) model (5) that explicitly takes into account the contemporaneous relationships of the variates that constitute $\mathbf{Y}_{*,i,t}$. The structural VAR(1) model that describes the gene expression data from the time-course experiment is:

$$\mathbf{Y}_{*,i,t} = \mathbf{B}\mathbf{Y}_{*,i,t} + \breve{\mathbf{A}}\mathbf{Y}_{*,i,t-1} + \breve{\boldsymbol{\varepsilon}}_{*,i,t}, \tag{7}$$

where $\breve{\mathbf{A}}$ a $p \times p$ coefficient matrix, $\mathbf{B}$ a $p \times p$ coefficient matrix with a zero diagonal such that $\mathbf{I}_{p\times p} - \mathbf{B}$ is non-singular, and $\breve{\boldsymbol{\varepsilon}}_{*,i,t}$ a $p \times 1$ vector with the errors. Throughout it is assumed that $\breve{\boldsymbol{\varepsilon}}_{*,i,t} \sim \mathcal{N}(\mathbf{0}_{p\times 1}, \breve{\boldsymbol{\Sigma}}_{\varepsilon})$ and $\mathrm{Cov}(\breve{\boldsymbol{\varepsilon}}_{*,i_1,t_1}, \breve{\boldsymbol{\varepsilon}}_{*,i_2,t_2}) = \mathbf{0}$ if $t_1 \neq t_2$ or $i_1 \neq i_2$. Thus, for a 3-gene pathway model (7) becomes:

$$\begin{cases} Y_{1,i,t} = b_{12}Y_{2,i,t-1} + b_{13}Y_{3,i,t} + \breve{a}_{11}Y_{1,i,t-1} + \breve{a}_{12}Y_{2,i,t-1} + \breve{a}_{13}Y_{3,i,t-1} + \breve{\varepsilon}_{1,i,t}, \\ Y_{2,i,t} = b_{21}Y_{1,i,t-1} + b_{23}Y_{3,i,t} + \breve{a}_{21}Y_{1,i,t-1} + \breve{a}_{22}Y_{2,i,t-1} + \breve{a}_{23}Y_{3,i,t-1} + \breve{\varepsilon}_{2,i,t}, \\ Y_{3,i,t} = b_{31}Y_{1,i,t-1} + b_{32}Y_{2,i,t} + \breve{a}_{31}Y_{1,i,t-1} + \breve{a}_{32}Y_{2,i,t-1} + \breve{a}_{33}Y_{3,i,t-1} + \breve{\varepsilon}_{3,i,t}, \end{cases}$$

where e.g. $Y_{1,i,t}$ does not appear in the first equations as $b_{11} = 0$.

The relations between the parameters of the canonical and structural VAR(1) models are:

$$\begin{aligned}
\boldsymbol{\varepsilon}_{*,i,t} &= (\mathbf{I}_{p\times p} - \mathbf{B})^{-1}\breve{\boldsymbol{\varepsilon}}_{*,i,t}, \\
\mathbf{A} &= (\mathbf{I}_{p\times p} - \mathbf{B})^{-1}\breve{\mathbf{A}}, \\
\boldsymbol{\Sigma}_{\varepsilon} &= (\mathbf{I}_{p\times p} - \mathbf{B})^{-1}\breve{\boldsymbol{\Sigma}}_{\varepsilon}[(\mathbf{I}_{p\times p} - \mathbf{B})^{-1}]^{\mathrm{T}},
\end{aligned}$$

as the canonical VAR(1) model (5) is obtained through pre-multiplication of the structural VAR(1) model (7) by $(\mathbf{I}_{p\times p} - \mathbf{B})^{-1}$.

8.3 Mean, variance, and autocovariance

Under the above assumptions the vector of gene expression levels of sample i at time point t follows a multivariate normal distribution: $\mathbf{Y}_{*,i,t} \sim \mathcal{N}(\boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y)$. The mean of $\mathbf{Y}_{*,i,t}$ is:

$$\begin{aligned}
E(\mathbf{Y}_{*,i,t} \,|\, \mathbf{Y}_{*,i,t-1}, \ldots, \mathbf{Y}_{*,i,1}) &= E(\boldsymbol{\nu}) + \mathbf{A}E\big(\mathbf{Y}_{*,i,t-1}\big) + E\big(\boldsymbol{\varepsilon}_{*,i,t}\big) \\
&= \boldsymbol{\nu} + \mathbf{A}E\big(\boldsymbol{\nu} + \mathbf{A}\mathbf{Y}_{*,i,t-2} + \boldsymbol{\varepsilon}_{*,i,t-1}\big) \\
&= \boldsymbol{\nu} + \mathbf{A}\boldsymbol{\nu} + \mathbf{A}^2E\big(\boldsymbol{\nu} + \mathbf{A}\mathbf{Y}_{*,i,t-3} + \boldsymbol{\varepsilon}_{*,i,t-2}\big) \\
&= \ldots \\
&= \big(\mathbf{I}_{p\times p} + \mathbf{A} + \mathbf{A}^2 + \ldots + \mathbf{A}^{\mathcal{T}-1}\big)\boldsymbol{\nu}.
\end{aligned}$$

If $\mathcal{T} \to \infty$, the mean $\boldsymbol{\mu}_y$ converges to $\big(\mathbf{I}_{p\times p} - \mathbf{A}\big)^{-1}\boldsymbol{\nu}$. This assumes that all eigenvalues of $\mathbf{A}$ are smaller than 1 in an absolute sense.

The variance of the process is a little bit more elaborate. Note that from the definition of the model (5) it is clear that $\boldsymbol{\Sigma}_y$ satisfies:

$$\boldsymbol{\Sigma}_y = \mathbf{A}\boldsymbol{\Sigma}_y\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_{\varepsilon}. \tag{8}$$

This equation is known as the discrete Lyapunov equation. An analytic solution (solving for $\boldsymbol{\Sigma}_y$) of this equation exists. To arrive at the solution apply the vec$(\cdot)$ operator, defined by

$$\mathrm{vec}(\mathbf{A}) = \mathrm{vec}\left(\begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1p} \\ a_{21} & a_{22} & \ldots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \ldots & a_{pp} \end{pmatrix}\right) = (a_{11}, \ldots, a_{p1}, a_{12}, \ldots, a_{p2}, \ldots, a_{1p}, \ldots, a_{pp})^{\mathrm{T}},$$

to both sides of Equation (8):

$$\begin{aligned}
\mathrm{vec}(\boldsymbol{\Sigma}_{\varepsilon}) &= \mathrm{vec}(\boldsymbol{\Sigma}_y) - \mathrm{vec}(\mathbf{A}\boldsymbol{\Sigma}_y\mathbf{A}^{\mathrm{T}}) \\
&= \mathrm{vec}(\boldsymbol{\Sigma}_y) - (\mathbf{A}\otimes\mathbf{A})\mathrm{vec}(\boldsymbol{\Sigma}_y) \\
&= (\mathbf{I}_{p^2\times p^2} - \mathbf{A}\otimes\mathbf{A})\mathrm{vec}(\boldsymbol{\Sigma}_y),
\end{aligned}$$

where the identity $\mathrm{vec}(\mathbf{ABC}) = (\mathbf{C}^{\mathrm{T}}\otimes\mathbf{A})\mathrm{vec}(\mathbf{B})$ is used. Thus:

$$\mathrm{vec}(\boldsymbol{\Sigma}_y) = (\mathbf{I}_{p^2\times p^2} - \mathbf{A}\otimes\mathbf{A})^{-1}\,\mathrm{vec}(\boldsymbol{\Sigma}_{\varepsilon}),$$

presuming the inverse exists. With respect to this existence note that *i)* the eigenvalues of $\mathbf{A}\otimes\mathbf{A}$ are given by $\lambda_{j_1}(\mathbf{A})\lambda_{j_2}(\mathbf{A})$ for $j_1, j_2 = 1, \ldots, p$, and *ii)* the eigenvalues of $\mathbf{I}_{p^2\times p^2} - \mathbf{A}\otimes\mathbf{A}$ are then of the form $1 - \lambda_{j_1}(\mathbf{A})\lambda_{j_2}(\mathbf{A})$. Consequently, the inverse (and thus the solution) above exists if for any eigenvalue of $\mathbf{A}$ its reciprocal is not an eigenvalue of $\mathbf{A}$.

A well-known alternative solution of the discrete Lyapunov equation (8) is:

$$\boldsymbol{\Sigma}_y \;=\; \sum_{k=0}^{\infty} \mathbf{A}^k \boldsymbol{\Sigma}_\varepsilon (\mathbf{A}^{\mathrm{T}})^k.$$

To see this is indeed a solution of the discrete Lyapunov equation, simply substitute it:

$$\begin{aligned}\sum_{k=0}^{\infty} \mathbf{A}^k \boldsymbol{\Sigma}_\varepsilon (\mathbf{A}^{\mathrm{T}})^k &= \mathbf{A}\Big(\sum_{k=0}^{\infty} \mathbf{A}^k \boldsymbol{\Sigma}_\varepsilon (\mathbf{A}^{\mathrm{T}})^k\Big)\mathbf{A}^{\mathrm{T}} + \boldsymbol{\Sigma}_\varepsilon \\ &= \sum_{k=1}^{\infty} \mathbf{A}^k \boldsymbol{\Sigma}_\varepsilon (\mathbf{A}^{\mathrm{T}})^k + \boldsymbol{\Sigma}_\varepsilon \;=\; \sum_{k=0}^{\infty} \mathbf{A}^k \boldsymbol{\Sigma}_\varepsilon (\mathbf{A}^{\mathrm{T}})^k.\end{aligned}$$

Furthermore, the positive-definiteness of $\boldsymbol{\Sigma}_\varepsilon$ warrants the positive-definiteness of $\boldsymbol{\Sigma}_y$ (this follows directly from iterative application of results from Harville 2008). **Quote proposition number from Harville.**

We now investigate the convergence of this series solution. Hereto diagonalize $\mathbf{A}$ by means of its eigen-decomposition $\mathbf{A} = \mathbf{V}_a \mathbf{D}_a \mathbf{V}_a^{-1}$. The discrete Lyapunov equation then becomes:

$$\boldsymbol{\Sigma}_y \;=\; \mathbf{V}_a \mathbf{D}_a \mathbf{V}_a^{-1} \boldsymbol{\Sigma}_y (\mathbf{V}_a^{-1})^{\mathrm{T}} \mathbf{D}_a \mathbf{V}_a^{\mathrm{T}} + \boldsymbol{\Sigma}_\varepsilon.$$

Or, reformulated:

$$\tilde{\boldsymbol{\Sigma}}_y \;=\; \mathbf{D}_a \tilde{\boldsymbol{\Sigma}}_y \mathbf{D}_a + \tilde{\boldsymbol{\Sigma}}_\varepsilon,$$

where $\tilde{\boldsymbol{\Sigma}}_y = \mathbf{V}_a^{-1} \boldsymbol{\Sigma}_y (\mathbf{V}_a^{-1})^{\mathrm{T}}$ and $\tilde{\boldsymbol{\Sigma}}_\varepsilon = \mathbf{V}_a^{-1} \boldsymbol{\Sigma}_\varepsilon (\mathbf{V}_a^{-1})^{\mathrm{T}}$. The solution of this reformulated discrete Lyapunov equation is:

$$\tilde{\boldsymbol{\Sigma}}_y \;=\; \sum_{k=0}^{\infty} \mathbf{D}_a^k \tilde{\boldsymbol{\Sigma}}_\varepsilon \mathbf{D}_a^k.$$

Or, in terms of the elements of the matrices:

$$\begin{aligned}(\tilde{\boldsymbol{\Sigma}}_y)_{j_1,j_2} &= \sum_{k=0}^{\infty} (\mathbf{D}_a^k \tilde{\boldsymbol{\Sigma}}_\varepsilon \mathbf{D}_a^k)_{j_1,j_2} \\ &= (\tilde{\boldsymbol{\Sigma}}_\varepsilon)_{j_1,j_2} \sum_{k=0}^{\infty} [(\mathbf{D}_a)_{j_1,j_1} (\mathbf{D}_a)_{j_2,j_2}]^k \\ &= \frac{(\tilde{\boldsymbol{\Sigma}}_\varepsilon)_{j_1,j_2}}{1 - (\mathbf{D}_a)_{j_1,j_1} (\mathbf{D}_a)_{j_2,j_2}},\end{aligned}$$

where we have used that $\mathbf{D}$ is diagonal. If one defines a $p \times p$ matrix $\mathbf{C}$ with $(\mathbf{C})_{j_1,j_2} = [1 - (\mathbf{D}_a)_{j_1,j_1} (\mathbf{D}_a)_{j_2,j_2}]^{-1}$, then:

$$\boldsymbol{\Sigma}_y \;=\; \mathbf{V}_a (\tilde{\boldsymbol{\Sigma}}_\varepsilon \circ \mathbf{C}) \mathbf{V}_a^{\mathrm{T}},$$

where the $\circ$ operator indicates the Hadamard product. From the above it should also be clear that $\boldsymbol{\Sigma}_y$ is well-defined by the series solution if $\lambda_{j_1}(\mathbf{A}) \neq [\lambda_{j_2}(\mathbf{A})]^{-1}$ for any couple j_1 and j_2.

Example 18. *Mean and variance of the feedforward motif.* Revisit Example 16. Clearly, as $\boldsymbol{\nu} = \mathbf{0}_{p\times 1}$, the mean $\boldsymbol{\mu}$ also equals $\mathbf{0}_{p\times 1}$. The variance of $\mathbf{Y}_{*,i,t}$ is:

$$\boldsymbol{\Sigma}_y \;=\; \begin{pmatrix} 0.2500 & 0.0000 & 0.0000 \\ 0.0000 & 1.8125 & -1.1250 \\ 0.0000 & -1.1250 & 5.1381 \end{pmatrix},$$

where numbers have been rounded.

Now we have defined the mean and variance of the time series process, the remaining interesting quantity in VAR(1) model is the autocovariance. The *autocovariance* is the covariance between the gene expression profile at time point t and that at some other time point. The autocovariance of the VAR(1) process is (for $\tau \geq 0$):

$$\begin{aligned}
\boldsymbol{\Gamma}(\tau) \;&=\; \mathrm{Cov}\big(\mathbf{Y}_{*,i,t}, \mathbf{Y}_{*,i,t+\tau}\big) \;=\; E\Big[\big(\mathbf{Y}_{*,i,t} - E(\mathbf{Y}_{*,i,t})\big)\big(\mathbf{Y}_{*,i,t+\tau} - E(\mathbf{Y}_{*,i,t+\tau})\big)^{\mathrm{T}}\Big] \\
&=\; E\Big[\Big(\mathbf{Y}_{*,i,t)} - E(\mathbf{Y}_{*,i,t})\Big)\Big(\mathbf{A}^{\tau}\mathbf{Y}_{*,i,t} + \sum_{k=0}^{\tau-1}\mathbf{A}^{k}\boldsymbol{\varepsilon}_{*,i,t+\tau-k} - E\big(\mathbf{A}^{\tau}\mathbf{Y}_{*,i,t}\big)\Big)^{\mathrm{T}}\Big] \\
&=\; E\Big[\big(\mathbf{Y}_{*,i,t} - E(\mathbf{Y}_{*,i,t})\big)\big(\mathbf{A}^{\tau}\mathbf{Y}_{*,i,t} - E(\mathbf{A}^{\tau}\mathbf{Y}_{*,i,t})\big)^{\mathrm{T}}\Big] \;=\; \mathrm{Cov}\big(\mathbf{Y}_{*,i,t}, \mathbf{Y}_{*,i,t}\big)[\mathbf{A}^{\tau}]^{\mathrm{T}} \\
&=\; \boldsymbol{\Sigma}_y[\mathbf{A}^{\tau}]^{\mathrm{T}}.
\end{aligned}$$

Similarly, for $\tau < 0$:

$$\boldsymbol{\Gamma}(\tau) \;=\; \mathrm{Cov}\big(\mathbf{Y}_{*,i,t-\tau}, \mathbf{Y}_{*,i,t}\big) \;=\; \mathbf{A}^{-\tau}\boldsymbol{\Sigma}_y.$$

In the above it is used that $\mathrm{Cov}\big(\boldsymbol{\varepsilon}_{*,i,t}, \boldsymbol{\varepsilon}_{*,i,t+\tau}\big) = 0$ if $\tau \neq 0$.

Associated with the autocovariance matrix are the *autocorrelation* and *crosscorrelation* (or sometimes *cross-autocorrelation*). Crosscorrelation is defined as $\mathrm{Cor}(Y_{j_1,i,t}, Y_{j_2,i,t+\tau})$ for $j_1 \neq j_2$ and $\tau \in \mathbb{N}$. The crosscorrelation is thus the correlation with lag τ between the time series of two different genes. The autocorrelation is reserved for the special case where $j_1 = j_2$. In the autocovariance matrix for the VAR(1) process, the diagonal elements are thus proportional to the autocorrelations and the off-diagonal elements to the crosscorrelations.

Finally, another frequently used quantity in times series is the *partial autocorrelation*. This is defined as:

$$\mathrm{Cor}(Y_{j,i,t}, Y_{j,i,t+\tau} \,|\, Y_{j,i,t+1}, \ldots, Y_{j,i,t+\tau-1}).$$

Hence, the correlation between the expression levels of a gene at two time points which is not propagated via the intermediate time points. The partial autocorrelation is used to determine the lag of the time series. As we assume the lag to equal one throughout, the partial autocovariance and partial autocorrelation appear to be of little interest here.

8.4 Estimation

8.4.1 Mean and autocovariance

The process mean is estimated by:

$$\hat{\boldsymbol{\mu}}_y \;=\; \frac{1}{n\mathcal{T}}\sum_{i=1}^{n}\sum_{t=1}^{\mathcal{T}}\mathbf{Y}_{*,i,t},$$

which has the following expectation:

$$E(\hat{\boldsymbol{\mu}}_y) \;=\; \frac{1}{n\mathcal{T}}\sum_{i=1}^{n}\sum_{t=1}^{\mathcal{T}} E(\mathbf{Y}_{*,i,t}) \;=\; \frac{1}{n\mathcal{T}}\sum_{i=1}^{n}\sum_{t=1}^{\mathcal{T}} \boldsymbol{\mu}_y \;=\; \boldsymbol{\mu}_y.$$

Hence, the estimator for the mean is unbiased.

A natural estimator for the autocovariance (confer Brockwell and Davis 2006) would be:

$$\hat{\boldsymbol{\Gamma}}(\tau) \;=\; \begin{cases} \frac{1}{n(\mathcal{T}-\tau)}\sum_{i=1}^{n}\sum_{t=1}^{\mathcal{T}-\tau}(\mathbf{Y}_{*,i,t}-\hat{\boldsymbol{\mu}}_y)\big(\mathbf{Y}_{*,i,t+\tau}-\hat{\boldsymbol{\mu}}_y\big)^{\mathrm{T}} & \text{for} \quad 0\le\tau\le\mathcal{T}-1, \\ \frac{1}{n(\mathcal{T}-\tau)}\sum_{i=1}^{n}\sum_{t=-\tau+1}^{\mathcal{T}}(\mathbf{Y}_{*,i,t}-\hat{\boldsymbol{\mu}}_y)\big(\mathbf{Y}_{*,i,t+\tau}-\hat{\boldsymbol{\mu}}_y\big)^{\mathrm{T}} & \text{for} \quad -\mathcal{T}+1\le\tau<0. \end{cases}$$

However, as we have centered the data, we consider the following estimator:

$$\hat{\boldsymbol{\Gamma}}(\tau) \;=\; \begin{cases} \frac{1}{n(\mathcal{T}-\tau)}\sum_{i=1}^{n}\sum_{t=1}^{\mathcal{T}-\tau}\mathbf{Y}_{*,i,t}\mathbf{Y}_{*,i,t+\tau}^{\mathrm{T}} & \text{for} \quad 0\le\tau\le\mathcal{T}-1, \\ \frac{1}{n(\mathcal{T}-\tau)}\sum_{i=1}^{n}\sum_{t=-\tau+1}^{\mathcal{T}}\mathbf{Y}_{*,i,t}\mathbf{Y}_{*,i,t+\tau}^{\mathrm{T}} & \text{for} \quad -\mathcal{T}+1\le\tau<0. \end{cases}$$

For the calculation of the expectation of the covariance estimator we thus assume that $\boldsymbol{\mu}_y = 0$ (as we have centered the gene expression levels around zero). Then, for $\tau \ge 0$, we have:

$$E[\hat{\boldsymbol{\Gamma}}(\tau)] \;=\; \frac{1}{n(\mathcal{T}-\tau)}\sum_{i=1}^{n}\sum_{t=1}^{\mathcal{T}-\tau} E[\mathbf{Y}_{*,i,t}\mathbf{Y}_{*,i,t+\tau}^{\mathrm{T}}] \;=\; \boldsymbol{\Gamma}(\tau).$$

Again, an unbiased estimator. Note that expectation of the 'natural' autocovariance estimator of Brockwell and Davis (2006) is not unbiased. It is however consistent.

Mark: When μ_y is assumed to be zero, the degrees of freedom also changes. Hence, divide by $n\mathcal{T}$ rather than $n(\mathcal{T}-\tau)$!

8.4.2 Model parameters A and $\boldsymbol{\Sigma}_\varepsilon$

We now turn to the estimation of the parameters of model (5). This is done by likelihood maximization. It should be noted that the ML estimators for $\mathbf{A}$ coincides with that of the least squares approach. As we have centered (per gene, within each sample) the expression values, we may consider $\boldsymbol{\nu} = \mathbf{0}$. Thus, $\mathbf{Y}_{*,i,t} \,|\, \mathbf{Y}_{*,i,t-1}, \ldots, \mathbf{Y}_{*,i,1} \sim \mathcal{N}(\mathbf{A}\mathbf{Y}_{*,i,t}, \boldsymbol{\Sigma}_\varepsilon)$. The likelihood is then:

$$\begin{aligned} & L(\mathbf{Y};\boldsymbol{\nu},\mathbf{A},\boldsymbol{\Sigma}_\varepsilon) \\ &= \prod_{i=1}^{n} P(\mathbf{Y}_{*,i,\mathcal{T}},\ldots,\mathbf{Y}_{*,i,1}) \;=\; \prod_{i=1}^{n}\prod_{t=1}^{\mathcal{T}} P(\mathbf{Y}_{*,i,t}\,|\,\mathbf{Y}_{*,i,t-1},\ldots,\mathbf{Y}_{*,i,1}) \\ &= \prod_{i=1}^{n}\prod_{t=2}^{\mathcal{T}} \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_\varepsilon|^{1/2}} \exp\big[-\big(\mathbf{Y}_{*,i,t}-\mathbf{A}\mathbf{Y}_{*,i,t-1}\big)^{\mathrm{T}}\boldsymbol{\Sigma}_\varepsilon^{-1}\big(\mathbf{Y}_{*,i,t}-\mathbf{A}\mathbf{Y}_{*,i,t-1}\big)\big] \\ &= \prod_{i=1}^{n}\prod_{t=2}^{\mathcal{T}} \frac{1}{(2\pi)^{p/2}|\boldsymbol{\Sigma}_\varepsilon|^{1/2}} \exp\big\{-\big[\mathbf{Y}_{*,i,t}-\big(\mathbf{Y}_{*,i,t-1}^{\mathrm{T}}\otimes\mathbf{I}_{p\times p}\big)\mathrm{vec}(\mathbf{A})\big]^{\mathrm{T}} \\ &\qquad\qquad \boldsymbol{\Sigma}_\varepsilon^{-1}\big[\mathbf{Y}_{*,i,t}-\big(\mathbf{Y}_{*,i,t-1}^{\mathrm{T}}\otimes\mathbf{I}_{p\times p}\big)\mathrm{vec}(\mathbf{A})\big]\big\} \end{aligned}$$

Equate the derivative of the log-likelihood with respect to $\mathrm{vec}(\mathbf{A})$ to zero:

$$0 \;=\; \sum_{i=1}^{n}\sum_{t=2}^{\mathcal{T}} \big(\mathbf{Y}_{*,i,t-1}^{\mathrm{T}}\otimes\mathbf{I}_{p\times p}\big)^{\mathrm{T}}\boldsymbol{\Sigma}_\varepsilon^{-1}\big(\mathbf{Y}_{*,i,t}-\big(\mathbf{Y}_{*,i,t-1}^{\mathrm{T}}\otimes\mathbf{I}_{p\times p}\big)\mathrm{vec}(\mathbf{A})\big).$$

After some algebraic manipulations this is rewritten to:

$$0 \;=\; \sum_{i=1}^{n}\sum_{t=2}^{\mathcal{T}} \big(\mathbf{Y}_{*,i,t-1} \otimes \mathbf{\Sigma}_{\varepsilon}^{-1}\big)\mathbf{Y}_{*,i,t} - \sum_{i=1}^{n}\sum_{t=1}^{\mathcal{T}} \big(\mathbf{Y}_{*,i,t-1}\mathbf{Y}_{*,i,t-1}^{\mathrm{T}} \otimes \mathbf{\Sigma}_{\varepsilon}^{-1}\big)\mathrm{vec}(\mathbf{A}).$$

Solve this estimation equation to obtain an estimate for $\mathbf{A}$:

$$\mathrm{vec}(\hat{\mathbf{A}}) \;=\; \Big(\sum_{i=1}^{n}\sum_{t=2}^{\mathcal{T}} \mathbf{Y}_{*,i,t-1}\mathbf{Y}_{*,i,t-1}^{\mathrm{T}} \otimes \mathbf{I}_{p\times p}\Big)^{-1} \sum_{i=1}^{n}\sum_{t=2}^{\mathcal{T}} \big(\mathbf{Y}_{*,i,t-1} \otimes \mathbf{I}_{p\times p}\big)\big(\mathbf{Y}_{*,i,t}\big).$$

Notice that this expression can be formulated differently:

$$\hat{\mathbf{A}} \;=\; \Big[\frac{1}{n(\mathcal{T}-1)}\sum_{i=1}^{n}\sum_{t=1}^{\mathcal{T}-1} \mathbf{Y}_{*,i,t+1}\mathbf{Y}_{*,i,t}^{\mathrm{T}}\Big]\Big[\frac{1}{n(\mathcal{T}-1)}\sum_{i=1}^{n}\sum_{t=1}^{\mathcal{T}-1} \mathbf{Y}_{*,i,t}\mathbf{Y}_{*,i,t}^{\mathrm{T}}\Big]^{-1}$$

This is a more convenient expression when dealing with high-dimensional data.

Finally, the estimator for $\mathbf{\Sigma}_{\varepsilon}$ is given by:

$$\begin{aligned}\hat{\mathbf{\Sigma}}_{\varepsilon} &= \frac{1}{n\mathcal{T}}\sum_{i=1}^{n}\sum_{t=2}^{\mathcal{T}} \big[\mathbf{Y}_{*,i,t} - \big(\mathbf{Y}_{*,i,t-1}^{\mathrm{T}} \otimes \mathbf{I}_{p\times p}\big)\mathrm{vec}(\mathbf{A})\big]\big[\mathbf{Y}_{*,i,t} - \big(\mathbf{Y}_{*,i,t-1}^{\mathrm{T}} \otimes \mathbf{I}_{p\times p}\big)\mathrm{vec}(\mathbf{A})\big]^{\mathrm{T}} \\ &= \frac{1}{n\mathcal{T}}\sum_{i=1}^{n}\sum_{t=2}^{\mathcal{T}} \big[\mathbf{Y}_{*,i,t} - \mathbf{A}\mathbf{Y}_{*,i,t-1}\big]\big[\mathbf{Y}_{*,i,t} - \mathbf{A}\mathbf{Y}_{*,i,t-1}\big]^{\mathrm{T}}.\end{aligned}$$

Thus, $\mathbf{\Sigma}_{\varepsilon}$ is simply estimated by the sample residual covariance matrix.

8.4.3 Bias of the OLS/ML estimator of A

For small $\mathcal{T}$, exactly the case we usually encounter, the OLS and ML estimator of $\mathbf{A}$, as derived above, are biased. An explicit expression of the bias of the 'natural estimator' of $\mathbf{A}$ in the VAR(1) model with $n = 1$ has been derived by Nicholls and Pope (1988). A sketch of their approach follows. Define:

$$\hat{\mathbf{\Gamma}}_{\mathcal{T}}(\tau) \;=\; \frac{1}{n(\mathcal{T}-\tau)}\sum_{i=1}^{n}\sum_{t=1}^{\mathcal{T}-\tau} \mathbf{Y}_{*,i,t}\mathbf{Y}_{*,i,t+\tau}^{\mathrm{T}},$$

where $\tau = -1$ or $\tau = 0$. Furthermore, define the random variables:

$$\begin{aligned}\mathbf{P}_{\mathcal{T}} &= \big[\hat{\mathbf{\Gamma}}_{\mathcal{T}}(-1) - \mathbf{\Gamma}(-1)\big][\mathbf{\Gamma}(0)]^{-1} \\ \mathbf{Q}_{\mathcal{T}} &= \big[\hat{\mathbf{\Gamma}}_{\mathcal{T}}(0) - \mathbf{\Gamma}(0)\big][\mathbf{\Gamma}(0)]^{-1} \;=\; \hat{\mathbf{\Gamma}}_{\mathcal{T}}(0)[\mathbf{\Gamma}(0)]^{-1} - \mathbf{I}_{p\times p}.\end{aligned}$$

The estimator of $\mathbf{A}$ can then be written as:

$$\begin{aligned}\hat{\mathbf{A}}_{\mathcal{T}} &= \hat{\mathbf{\Gamma}}_{\mathcal{T}}(-1)\big[\hat{\mathbf{\Gamma}}_{\mathcal{T}}(0)\big]^{-1} \\ &= \hat{\mathbf{\Gamma}}_{\mathcal{T}}(-1)\big[\hat{\mathbf{\Gamma}}_{\mathcal{T}}(0)\big]^{-1} + \mathbf{A}\mathbf{\Gamma}(0)[\hat{\mathbf{\Gamma}}_{\mathcal{T}}(0)]^{-1} - \mathbf{\Gamma}(-1)[\hat{\mathbf{\Gamma}}_{\mathcal{T}}(0)]^{-1} \\ &= \hat{\mathbf{\Gamma}}_{\mathcal{T}}(-1)[\mathbf{\Gamma}(0)]^{-1}\mathbf{\Gamma}(0)\big[\hat{\mathbf{\Gamma}}_{\mathcal{T}}(0)\big]^{-1} + \mathbf{A}(\mathbf{I}_{p\times p} + \mathbf{Q}_{\mathcal{T}})^{-1} - \mathbf{\Gamma}(-1)[\mathbf{\Gamma}(0)]^{-1}\mathbf{\Gamma}(0)[\hat{\mathbf{\Gamma}}_{\mathcal{T}}(0)]^{-1} \\ &= \{\hat{\mathbf{\Gamma}}_{\mathcal{T}}(-1)[\mathbf{\Gamma}(0)]^{-1} - \mathbf{\Gamma}(-1)[\mathbf{\Gamma}(0)]^{-1}\}\mathbf{\Gamma}(0)\big[\hat{\mathbf{\Gamma}}_{\mathcal{T}}(0)\big]^{-1} + \mathbf{A}(\mathbf{I}_{p\times p} + \mathbf{Q}_{\mathcal{T}})^{-1} \\ &= (\mathbf{A} + \mathbf{P}_{\mathcal{T}})(\mathbf{I}_{p\times p} + \mathbf{Q}_{\mathcal{T}})^{-1} \\ &= (\mathbf{A} + \mathbf{P}_{\mathcal{T}})\sum_{k=0}^{\infty}(-1)^k\mathbf{Q}_{\mathcal{T}}^{k} \\ &= \mathbf{A} + \mathbf{P}_{\mathcal{T}} - \mathbf{A}\mathbf{Q}_{\mathcal{T}} - \mathbf{P}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}} + \mathbf{A}\mathbf{Q}_{\mathcal{T}}^{2} + \mathbf{P}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}^{2} + \ldots\end{aligned}$$

Now take the expectation on both sides:

$$\begin{aligned} E(\hat{\mathbf{A}}_{\mathcal{T}}) &= \mathbf{A} + E(\mathbf{P}_{\mathcal{T}}) - E(\mathbf{A}\mathbf{Q}_{\mathcal{T}}) - E(\mathbf{P}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}) + E(\mathbf{A}\mathbf{Q}_{\mathcal{T}}^2) + E(\mathbf{P}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}^2) + \dots \\ &= \mathbf{A} + E(\mathbf{P}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}) + E(\mathbf{A}\mathbf{Q}_{\mathcal{T}}^2) + E(\mathbf{P}_{\mathcal{T}}\mathbf{Q}_{\mathcal{T}}^2) + \dots \end{aligned}$$

The remaining expectation are not (necessarily) zero, hence the OLS / ML estimator of $\mathbf{A}$ is biased (Nicholls and Pope, 1988). However, it is consistent (Lütkepohl, 2005).

8.4.4 Estimation with constraints on A

Add constraints. In particular, the sign of the coefficients may be known. Hence, work out details.

8.4.5 Ridge estimation

The ridge estimator for $\mathbf{A}$ is readily obtained by subtracting the term $\frac{1}{2}\mathrm{tr}[\mathrm{vec}(\mathbf{A})^{\mathrm{T}}\mathbf{\Lambda}\mathrm{vec}(\mathbf{A})]$ from the log-likelihood. In case $\mathbf{\Lambda} = \lambda\mathbf{I}_{p^2\times p^2}$ the penalty reduces to $\frac{\lambda}{2}\mathrm{tr}[\mathrm{vec}(\mathbf{A})^{\mathrm{T}}\mathbf{\Lambda}\mathrm{vec}(\mathbf{A})] = \frac{\lambda}{2}\|\mathbf{A}\|_2^2 = \frac{\lambda}{2}\sum_{j_1,j_2=1}^{p}(\mathbf{A})_{j_1,j_2}^2$. The ridge estimator thus becomes:

$$\mathrm{vec}[\hat{\mathbf{A}}(\lambda)] = \Big(\lambda\mathbf{I}_{p^2\times p^2} + \sum_{i=1}^{n}\sum_{t=2}^{\mathcal{T}}\mathbf{Y}_{*,i,t-1}\mathbf{Y}_{*,i,t-1}^{\mathrm{T}} \otimes \mathbf{I}_{p\times p}\Big)^{-1}\sum_{i=1}^{n}\sum_{t=2}^{\mathcal{T}}\big(\mathbf{Y}_{*,i,t-1}\otimes\mathbf{I}_{p\times p}\big)\big(\mathbf{Y}_{*,i,t}-\boldsymbol{\nu}\big).$$

Or, when we use the more efficient formulation of the estimator if $\mathbf{A}$ and write $\mathbf{\Lambda} = \lambda\mathbf{I}_{p\times p}$:

$$\hat{\mathbf{A}}(\lambda) = \Big(\frac{1}{n(\mathcal{T}-1)}\sum_{i=1}^{n}\sum_{t=1}^{\mathcal{T}-1}\mathbf{Y}_{*,i,t+1}\mathbf{Y}_{*,i,t}^{\mathrm{T}}\Big)\Big(\lambda\mathbf{I}_{p\times p} + \frac{1}{n(\mathcal{T}-1)}\sum_{i=1}^{n}\sum_{t=1}^{\mathcal{T}-1}\mathbf{Y}_{*,i,t}\mathbf{Y}_{*,i,t}^{\mathrm{T}}\Big)^{-1}. \quad (9)$$

In any case, ridge estimation in the VAR(1) model is analogous to that in the standard linear regression model.

Let us try to understand the effect of ridge penalty on the estimation of $\mathbf{A}$ within the VAR(1) model. As the ridge estimator is the product of a lag one autocovariance estimate and the inverse of a biased covariance estimat, consider the related object:

$$\begin{aligned} \mathbf{A}(\lambda) &= \mathbf{\Gamma}(-1)[\mathbf{\Gamma}(0) + \lambda\mathbf{I}_{p\times p}]^{-1} \\ &= \mathbf{A}\mathbf{\Sigma}_y(\mathbf{\Sigma}_y + \lambda\mathbf{I}_{p\times p})^{-1} \end{aligned}$$

Define a new stochastic process by convoluting the VAR(1) process with white noise:

$$\mathbf{Z}_{*,i,t} = \mathbf{Y}_{*,i,t} + \boldsymbol{\delta}_{*,i,t}, \quad (10)$$

with $\mathbf{Y}_{*,i,t}$ as in the original process (5) and $\boldsymbol{\delta}_{*,i,t} \sim \mathcal{N}(\mathbf{0}_{p\times 1}, \lambda\mathbf{I}_{p\times p})$. The variance of this process is then: $\mathbf{\Sigma}_y + \lambda\mathbf{I}_{p\times p}$ and $\mathbf{Z}_{*,i,t}$ has the same autocovariance as $\mathbf{Y}_{*,i,t}$. $\mathbf{A}(\lambda)$ is then the product of the autocovariance with lag -1 and inverse of the variance the subprocess $\mathbf{Z}_{*,i,t}$. Inclusion of the ridge penalty could be thought of as adding noise to the process.

Note that this 'adding noise' to the process is *not* equivalent to replacing $\mathbf{\Sigma}_\varepsilon$ by $\mathbf{\Sigma}_\varepsilon + \lambda\mathbf{I}_{p\times p}$. This would amount to redefining the VAR(1) model to:

$$\mathbf{Y}_{*,i,t} = \mathbf{A}\mathbf{Y}_{*,i,t-1} + \varepsilon_{*,i,t} + \boldsymbol{\delta}_{*,i,t}.$$

In this model the $\boldsymbol{\delta}_{*,i,t}$ are part of the innovations and propagated, which is not the case in the convolution. Among others this yields a covariance matrix unequal to $\mathbf{\Sigma}_y + \lambda\mathbf{I}_{p\times p}$. To see this consider the discrete Lyapunov equation modified in accordance with this model:

$$\mathbf{\Sigma}_y(\lambda) = \mathbf{A}\mathbf{\Sigma}_y(\lambda)\mathbf{A}^{\mathrm{T}} + \mathbf{\Sigma}_\varepsilon + \lambda\mathbf{I}_{p\times p}.$$

This thus implies that

$$\begin{aligned}\boldsymbol{\Sigma}_y(\lambda) &= \sum_{k=0}^{\infty}\mathbf{A}^k[\boldsymbol{\Sigma}_\varepsilon+\lambda\mathbf{I}_{p\times p}](\mathbf{A}^{\mathrm{T}})^k \\ &= \sum_{k=0}^{\infty}\mathbf{A}^k\boldsymbol{\Sigma}_\varepsilon(\mathbf{A}^{\mathrm{T}})^k+\lambda\sum_{k=0}^{\infty}\mathbf{A}^k(\mathbf{A}^{\mathrm{T}})^k \\ &= \boldsymbol{\Sigma}_y+\lambda\mathbf{I}_{p\times p}+\lambda\sum_{k=1}^{\infty}\mathbf{A}^k(\mathbf{A}^{\mathrm{T}})^k,\end{aligned}$$

which is unequal to $\boldsymbol{\Sigma}_y+\lambda\mathbf{I}_{p\times p}$.

Alternatively, define the

$$\begin{pmatrix}\mathbf{Z}_{*,i,t}\\ \mathbf{u}_{*,i,t}\end{pmatrix} = \begin{pmatrix}\mathbf{Y}_{*,i,t}+\boldsymbol{\delta}_{*,i,t}\\ \mathbf{A}\,\boldsymbol{\delta}_{*,i,t}\end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix}\mathbf{0}_{p\times 1}\\ \mathbf{0}_{p\times 1}\end{pmatrix},\begin{pmatrix}\boldsymbol{\Sigma}_y+\lambda\mathbf{I}_{p\times p} & \lambda\mathbf{A}\\ \lambda\mathbf{A}^{\mathrm{T}} & \lambda\mathbf{A}\mathbf{A}^{\mathrm{T}}\end{pmatrix}\right).$$

Then:

$$\mathrm{Var}(\mathbf{u}_{*,i,t}\,|\,\mathbf{Z}_{*,i,t}) = \mathbf{A}\mathbf{A}^{\mathrm{T}}-\lambda\mathbf{A}(\boldsymbol{\Sigma}_y+\lambda\mathbf{I}_{p\times p})^{-1}\mathbf{A}^{\mathrm{T}} = \mathbf{A}(\lambda)\mathbf{A}^{\mathrm{T}}.$$

The last equality follows from:

$$\begin{aligned}\mathbf{A}(\lambda) &= \boldsymbol{\Gamma}(-1)[\boldsymbol{\Gamma}(0)+\lambda\mathbf{I}_{p\times p}]^{-1} \\ &= \mathbf{A}\boldsymbol{\Sigma}_y(\boldsymbol{\Sigma}_y+\lambda\mathbf{I}_{p\times p})^{-1} \\ &= \mathbf{A}(\boldsymbol{\Sigma}_y+\lambda\mathbf{I}_{p\times p})(\boldsymbol{\Sigma}_y+\lambda\mathbf{I}_{p\times p})^{-1}-\lambda\mathbf{A}(\boldsymbol{\Sigma}_y+\lambda\mathbf{I}_{p\times p})^{-1} \\ &= \mathbf{A}-\lambda\mathbf{A}(\boldsymbol{\Sigma}_y+\lambda\mathbf{I}_{p\times p})^{-1}.\end{aligned}$$

Post-multiplying with $\mathbf{A}^{\mathrm{T}}$ gives the desired equality.**What is the point?**

The above also suggests that $\mathbf{A}(\lambda)\mathbf{A}^{\mathrm{T}}$ is symmetric. To see whether this is the case, note that:

$$\begin{aligned}\mathbf{A}(\lambda)\boldsymbol{\Sigma}_y\mathbf{A}^{\mathrm{T}}-\mathbf{A}\boldsymbol{\Sigma}_y\mathbf{A}^{\mathrm{T}} &= \mathbf{A}[\boldsymbol{\Sigma}_y(\boldsymbol{\Sigma}_y+\lambda\mathbf{I}_{p\times p})^{-1}\boldsymbol{\Sigma}_y-\boldsymbol{\Sigma}_y]\mathbf{A}^{\mathrm{T}} \\ &= \mathbf{A}[\lambda(\boldsymbol{\Sigma}_y+\lambda\mathbf{I}_{p\times p})^{-1}\boldsymbol{\Sigma}_y]\mathbf{A}^{\mathrm{T}} \\ &= \lambda[\mathbf{A}(\lambda)\mathbf{A}^{\mathrm{T}}]^{\mathrm{T}}.\end{aligned}$$

Then, substition of this results and the previous yields:

$$\lambda\mathbf{A}(\lambda)\mathbf{A}^{\mathrm{T}}-\lambda[\mathbf{A}(\lambda)\mathbf{A}^{\mathrm{T}}]^{\mathrm{T}} = \mathbf{0}.$$

Indeed, $\mathbf{A}(\lambda)\mathbf{A}^{\mathrm{T}}$ is symmetric.

ADD RIDGE ESTIMATION OF $\boldsymbol{\Sigma}_\varepsilon$

8.5 Granger causality

Granger causality is a concept that refers to whether one variable is useful in predicting another. It exploits the 'arrow of time' logic that a cause has to precede the effect. Pursuing this reasoning, if one variable affects another, knowledge of the former should improve the prediction of the latter.

Definition 1. Let $\mathbf{Y}_{*,i,t}$ follow a VAR(1) process. The random variable $Y_{j_1,i,t}$ is Granger noncausal in the mean (does not Granger cause in the mean) the random variable $Y_{j_2,i,t}$ if

$$E(Y_{j_2,i,t} \,|\, \mathbf{Y}_{*,i,t-1}, \mathbf{Y}_{*,i,t-2}, \ldots) \quad = \quad E(Y_{j_2,i,t} \,|\, \mathbf{Y}_{*\setminus j_1,i,t-1}, \mathbf{Y}_{*\setminus j1,i,t-2}, \ldots).$$

That is, the conditional expectation (forecast) of $Y_{j_2,i,t}$ does not improve by inclusion of past information of the $Y_{j_1,i,t}$.

Analogous definitions (e.g. Granger causality in distribution) replace the conditional expectation by the conditional variance, MSE, or distribution.

For illustration consider the bivariate VAR(1) process defined by:

$$\left\{ \begin{array}{rcl} Y_{1,i,t} & = & \nu_1 + a_{11} Y_{1,i,t-1} + a_{12} Y_{2,i,t-1} + \varepsilon_{1,i,t}, \\ Y_{2,i,t} & = & \nu_2 + a_{21} Y_{1,i,t-1} + a_{22} Y_{2,i,t-1} + \varepsilon_{2,i,t}, \end{array} \right.$$

If $a_{21} = 0$, then $Y_{1,i,t}$ does not Granger cause $Y_{2,i,t}$. If in addition $a_{12} \neq 0$, then $Y_{2,i,t}$ Granger causes $Y_{1,i,t}$. Hence, Granger causality is not a commutative property.

Finally, it is important to stress that Granger causality is not true causality. For instance, suppose the true causal structure is given by $Y_{1,i,t} \leftarrow Y_{2,i,t} \rightarrow Y_{3,i,t}$, where $Y_{2,i,t}$ is not observed. Then, usually one concludes that $Y_{1,i,t}$ Granger causes $Y_{3,i,t}$, and vice versa. Hence, indirect effects are not accounted for by Granger causality.

8.6 VAR(1) and graphical modelling

We now focus on the relation between the parameters of the VAR(1) model and conditional independence relationships among the variates of the model.

8.6.1 Contemporaneous conditional dependencies

First we address the conditional independency relations between the variates of $\mathbf{Y}_{*,i,t}$, which are called *contemporaneous CI relations*. Hereto consider the structural VAR(1) model (7). The conditional distribution of $\mathbf{Y}_{*,i,t}$ given $\mathbf{Y}_{*,i,t-1}$ is then:

$$\mathbf{Y}_{*,i,t} \,|\, \mathbf{Y}_{*,i,t-1} \quad \sim \quad \mathcal{N}\{\mathbf{A}\mathbf{Y}_{*,i,t-1}, (\mathbf{I}_{p\times p} - \mathbf{B})^{-1} \breve{\mathbf{\Sigma}}_{\varepsilon} [(\mathbf{I}_{p\times p} - \mathbf{B})^{-1}]^{\mathrm{T}}\}.$$

Using stuff to prove the well-known but horribly formulated lemma, we can link the elements of B to the partial correlations. The conditional independences between nodes at the same time point may thus be obtained from inversion of $\mathbf{\Sigma}_{\varepsilon}$, as in the regular Gaussian graphical model.

A second route is provided by Lemma 6.1.2 of Reale (1998), and reformulated here:

Lemma 2. Let $\mathbf{Y}_{*,i,t}$ follow a multivariate time series model, e.g. VAR(1), with canonical innovations $\boldsymbol{\varepsilon}_{*,i,t}$. Then:

$$\begin{array}{rcl} & & \mathrm{Corr}(\varepsilon_{j_1,i,t}, \varepsilon_{j_2,i,t} \,|\, \boldsymbol{\varepsilon}_{*,i,t}, \boldsymbol{\varepsilon}_{*,i,t-1}, \ldots \setminus \{\varepsilon_{j_1,i,t}, \varepsilon_{j_2,i,t}\}) \\ & = & \mathrm{Corr}(\varepsilon_{j_1,i,t}, \varepsilon_{j_2,i,t} \,|\, \mathbf{Y}_{*,i,t}, \mathbf{Y}_{*,i,t-1}, \ldots \setminus \{\varepsilon_{j_1,i,t}, \varepsilon_{j_2,i,t}\}) \\ & = & \mathrm{Corr}(Y_{j_1,i,t}, Y_{j_2,i,t} \,|\, \mathbf{Y}_{*,i,t}, \mathbf{Y}_{*,i,t-1}, \ldots \setminus \{\varepsilon_{j_1,i,t}, \varepsilon_{j_2,i,t}\}), \end{array}$$

for $j_1, j_2 \in \{1, \ldots, p\}$.

As a direct consequence of this lemma Reale (1998) deduces:

Corollary 4. Let $\mathcal{G}$ be the moral graph (representing the conditional independence relations) of $\mathbf{Y}_{*,i,t}, \mathbf{Y}_{*,i,t-1}, \ldots$, and $\mathcal{G}_t$ the subgraph of $\mathcal{G}$ related only to $\mathbf{Y}_{*,i,t}$ and its conditional independencies. The $\mathcal{G}_t$ is identical to the moral graph of $\boldsymbol{\varepsilon}_{*,i,t}$.

Alternatively, consider the covariance between $\mathbf{Y}_{*,i,t-1}$ and $\mathbf{Y}_{*,i,t-1}$:

$$\operatorname{Cov}\begin{pmatrix} \mathbf{Y}_{*,i,t-1} \\ \mathbf{Y}_{*,i,t} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_y & \boldsymbol{\Sigma}_y\mathbf{A}^{\mathbf{T}} \\ \mathbf{A}\boldsymbol{\Sigma}_y & \boldsymbol{\Sigma}_y \end{pmatrix}.$$

Using the analytic expression for the inverse of a 2×2 block matrix, the corresponding precision matrix equals:

$$\begin{pmatrix} \boldsymbol{\Sigma}_y^{-1} + \mathbf{A}^{\mathrm{T}}(\boldsymbol{\Sigma}_y - \mathbf{A}\boldsymbol{\Sigma}_y\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A} & -\mathbf{A}^{\mathrm{T}}(\boldsymbol{\Sigma}_y - \mathbf{A}\boldsymbol{\Sigma}_y\mathbf{A}^{\mathrm{T}})^{-1} \\ -(\boldsymbol{\Sigma}_y - \mathbf{A}\boldsymbol{\Sigma}_y\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A} & (\boldsymbol{\Sigma}_y - \mathbf{A}\boldsymbol{\Sigma}_y\mathbf{A}^{\mathrm{T}})^{-1} \end{pmatrix}.$$

As $\boldsymbol{\Sigma}_y$, $\boldsymbol{\Sigma}_\varepsilon$ and $\mathbf{A}$ satisfy the Lyapunov equation, we may write $\boldsymbol{\Sigma}_y - \mathbf{A}\boldsymbol{\Sigma}_y\mathbf{A}^{\mathrm{T}} = \boldsymbol{\Sigma}_\varepsilon$. Hence, the precision matrix simplifies to:

$$\begin{pmatrix} \boldsymbol{\Sigma}_y^{-1} + \mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{A} & -\mathbf{A}^{\mathrm{T}}\boldsymbol{\Sigma}_\varepsilon^{-1} \\ -\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{A} & \boldsymbol{\Sigma}_\varepsilon^{-1} \end{pmatrix}.$$

The bottom-right block of the precision matrix implies that the conditional independencies within $\mathbf{Y}_{*,i,t}$ can be read from $\boldsymbol{\Sigma}_\varepsilon^{-1}$. Moreover, it suggests that those between $\mathbf{Y}_{*,i,t-1}$ and $\mathbf{Y}_{*,i,t}$ can be deduced from $\boldsymbol{\Sigma}_\varepsilon^{-1}\mathbf{A}$.

Let $\mathbf{Y}_{*,i,t}$ follow a VAR(1) process. Then: $\mathbf{Y}_{*,i,t} \,|\, \mathbf{Y}_{*,i,t-1}, \mathbf{Y}_{*,i,t-2}, \ldots = \mathbf{Y}_{*,i,t} \,|\, \mathbf{Y}_{*,i,t-1} \sim \mathcal{N}(\mathbf{A}\mathbf{Y}_{*,i,t-1}, \boldsymbol{\Sigma}_\varepsilon)$. The density of $\mathbf{Y}_{*,i,t} \,|\, \mathbf{Y}_{*,i,t-1}$ can then be written as:

$$\begin{aligned} & P(\mathbf{Y}_{*,i,t} \,|\, \mathbf{Y}_{*,i,t-1}) \\ & \quad \propto \exp\left[-(\mathbf{Y}_{*,i,t} - \mathbf{A}\mathbf{Y}_{*,i,t-1})^{\mathrm{T}}\, \boldsymbol{\Sigma}_\varepsilon^{-1}\, (\mathbf{Y}_{*,i,t} - \mathbf{A}\mathbf{Y}_{*,i,t-1})/2\right] \\ & \quad = \exp\Big[-\frac{1}{2}\sum_{j_1,j_2=1}^{p} (\mathbf{Y}_{*,i,t} - \mathbf{A}\mathbf{Y}_{*,i,t-1})_{j_1,1}\, (\boldsymbol{\Sigma}_\varepsilon^{-1})_{j_1,j_2}\, (\mathbf{Y}_{*,i,t} - \mathbf{A}\mathbf{Y}_{*,i,t-1})_{j_2,1}\Big] \\ & \quad = \exp\Big[-\frac{1}{2}\sum_{j_1,j_2=1}^{p} (\boldsymbol{\Sigma}_\varepsilon^{-1})_{j_1,j_2}\, Y_{j_1,i,t}\, Y_{j_2,i,t} + (\boldsymbol{\Sigma}_\varepsilon^{-1})_{j_1,j_2}\, (\mathbf{A}\mathbf{Y}_{*,i,t-1})_{j_1,1}\, (\mathbf{A}\mathbf{Y}_{*,i,t-1})_{j_2,1} \\ & \qquad\qquad -(\boldsymbol{\Sigma}_\varepsilon^{-1})_{j_1,j_2}\, Y_{j_1,i,t}\, (\mathbf{A}\mathbf{Y}_{*,i,t-1})_{j_2,1} - (\boldsymbol{\Sigma}_\varepsilon^{-1})_{j_1,j_2}\, Y_{j_2,i,t}\, (\mathbf{A}\mathbf{Y}_{*,i,t-1})_{j_1,1}\Big]. \end{aligned}$$

If $(\boldsymbol{\Sigma}_\varepsilon^{-1})_{j_1,j_2} = 0$ for some $j_1, j_2 \in \{1, \ldots, p\}$, it is clear that the corresponding term $Y_{j_1,i,t}Y_{j_2,i,t}$ does not appear in the conditional density above. Hence, the conditional density can be factorized such that $Y_{j_1,i,t}$ and $Y_{j_2,i,t}$ do not occur in the same factor.

Example 19. *Illustration of conditional independence of contemporaneous variates.*

8.6.2 Conditional dependencies across time

How does this relate to a DAG?

8.7 Inference on parameters

How to infer whether elements of $\mathbf{A}$ and (inverse of) $\boldsymbol{\Sigma}_\varepsilon$ are zero?

8.8 Illustration

Relate to do-calculus of Pearl? Is this related to impulse response analysis? Affect one variable, see what happens downstream? How could the VAR(1) machinery presented here benefit cancer research?

9 Two molecular levels

9.1 The VARX(1) model

The gene expression data from the time-course experiment is modeled by a VAR(1) (first-order vector autoregressive) process with time-varying vector of covariates:

$$\mathbf{Y}_{*,i,t} \quad = \quad \boldsymbol{\nu} + \mathbf{A}\mathbf{Y}_{*,i,t-1} + \mathbf{C}\mathbf{X}_{*,i,t} + \boldsymbol{\varepsilon}_{*,i,t}, \tag{11}$$

where $\boldsymbol{\nu}$ the $p \times 1$ intercept vector, $\mathbf{A}$ a $p \times p$ coefficient matrix, $\mathbf{C}$ a $p \times q$ coefficient matrix, and $\boldsymbol{\varepsilon}_{*,i,t+1}$ a $p \times 1$ vector with the errors. Throughout it is assumed that $\boldsymbol{\varepsilon}_{*,i,t} \sim \mathcal{N}(\mathbf{0}_{p\times 1}, \boldsymbol{\Sigma}_\varepsilon)$ and $\mathrm{Cov}(\boldsymbol{\varepsilon}_{*,i_1,t_1}, \boldsymbol{\varepsilon}_{*,i_2,t_2}) = \mathbf{0}$ if $t_1 \neq t_2$ or $i_1 \neq i_2$.

When the covariates represent the expression of microRNAs, it would perhaps be more realistic to consider a model in which the microRNAs also follow a VAR(1) process. Model (11) thus becomes:

$$\left\{ \begin{array}{rclcccccc} \mathbf{Y}_{*,i,t} & = & \boldsymbol{\nu} & + & \mathbf{A}\mathbf{Y}_{*,i,t-1} & + & \mathbf{C}\mathbf{X}_{*,i,t} & + & \boldsymbol{\varepsilon}^{(y)}_{*,i,t} \\ \mathbf{X}_{*,i,t} & = & \boldsymbol{\gamma} & + & & + & \mathbf{D}\mathbf{X}_{*,i,t} & + & \boldsymbol{\varepsilon}^{(x)}_{*,i,t} \end{array} \right.$$

This is similar to Carel's stuff.

9.2 Parameter estimation

As gene expression levels are assumed to be gene-wise centered around zero, it seems reasonable to set $\boldsymbol{\nu} = \mathbf{0}_{p\times p}$. Model (11) reduces to:

$$\mathbf{Y}_{*,i,t} \quad = \quad \mathbf{A}\mathbf{Y}_{*,i,t-1} + \mathbf{C}\mathbf{X}_{*,i,t} + \boldsymbol{\varepsilon}_{*,i,t}.$$

This may be rewritten as:

$$\begin{aligned} \mathbf{Y}_{*,i,t} &= (\mathbf{A} \,|\, \mathbf{C}) \begin{pmatrix} \mathbf{Y}_{*,i,t-1} \\ \mathbf{X}_{*,i,t} \end{pmatrix} + \boldsymbol{\varepsilon}_{*,i,t} \\ &:= (\mathbf{A} \,|\, \mathbf{C})\, \mathbf{Z} + \boldsymbol{\varepsilon}_{*,i,t}, \end{aligned}$$

where $(\mathbf{A} \,|\, \mathbf{C})$ is a $p \times (p+q)$ dimensional matrix and $\mathbf{Z}$ a $(p+q) \times 1$ dimensional vector. Vectorizing (that is, applying the $\mathrm{vec}(\cdot)$ operator) to both sides of the equation, one obtains:

$$\begin{aligned} \mathbf{Y}_{*,i,t} &= \mathrm{vec}[(\mathbf{A} \,|\, \mathbf{C})\, \mathbf{Z}] + \boldsymbol{\varepsilon}_{*,i,t} \\ &= (\mathbf{Z} \otimes \mathbf{I}_{p\times p})\, \mathrm{vec}[(\mathbf{A} \,|\, \mathbf{C})] + \boldsymbol{\varepsilon}_{*,i,t} \\ &= [(\mathbf{Y}^{\mathrm{T}}_{*,i,t-1} \,|\, \mathbf{X}^{\mathrm{T}}_{*,i,t}) \otimes \mathbf{I}_{p\times p}]\, \mathrm{vec}[(\mathbf{A} \,|\, \mathbf{C})] + \boldsymbol{\varepsilon}_{*,i,t}, \end{aligned}$$

where $[(\mathbf{Y}^{\mathrm{T}}_{*,i,t-1} \,|\, \mathbf{X}^{\mathrm{T}}_{*,i,t}) \otimes \mathbf{I}_{p\times p}]$ and $\mathrm{vec}[(\mathbf{A} \,|\, \mathbf{C})]$ are of dimensions $p \times (p^2 + pq)$ and $(p^2 + pq) \times 1$. The product of these terms has the same dimension as $Y_{*,i,t}$: $p \times 1$ as desired.

The last formulation of model (11) is in the form of a standard linear regression model and facilitates estimation. The log-likelihood based loss function for $\mathrm{vec}[(\mathbf{A} \,|\, \mathbf{C})]$ is:

$$\sum_{i=1}^{n} \sum_{t=2}^{\mathcal{T}} \left\{ \mathbf{Y}_{*,i,t} - [(\mathbf{Y}^{\mathrm{T}}_{*,i,t-1} \,|\, \mathbf{X}^{\mathrm{T}}_{*,i,t}) \otimes \mathbf{I}_{p\times p}]\, \boldsymbol{\beta} \right\}^{\mathrm{T}} \boldsymbol{\Sigma}_\varepsilon^{-1} \left\{ \mathbf{Y}_{*,i,t} - [(\mathbf{Y}^{\mathrm{T}}_{*,i,t-1} \,|\, \mathbf{X}^{\mathrm{T}}_{*,i,t}) \otimes \mathbf{I}_{p\times p}]\, \boldsymbol{\beta} \right\},$$

where $\boldsymbol{\beta} = \mathrm{vec}[(\mathbf{A} \,|\, \mathbf{C})]$. Differentiate this loss function with respect to $\boldsymbol{\beta}$, equate the resulting derivate to zero, and solve for $\boldsymbol{\beta}$ to arrive at the OLS estimator. These steps are algebraic analogous

to those in the estimation of the parameters of the regular VAR(1) model (5). The ML estimator (which coincides with the OLS estimator) is thus:

$$\hat{\boldsymbol{\beta}} = \left(\sum_{i=1}^{n} \sum_{t=2}^{\mathcal{T}} \begin{pmatrix} \mathbf{Y}_{*,i,t-1} \\ \mathbf{X}_{*,i,t} \end{pmatrix} (\mathbf{Y}_{*,i,t-1}^{\mathrm{T}} \,|\, \mathbf{X}_{*,i,t}^{\mathrm{T}}) \otimes \mathbf{I}_{p\times p} \right)^{-1} \sum_{i=1}^{n} \sum_{t=2}^{\mathcal{T}} \left(\begin{pmatrix} \mathbf{Y}_{*,i,t-1} \\ \mathbf{X}_{*,i,t} \end{pmatrix} \otimes \mathbf{I}_{p\times p} \right) \mathbf{Y}_{*,i,t}.$$

Notice that this expression can be formulated differently **to do**:

$$\hat{\boldsymbol{\beta}} = \Big[\frac{1}{n(\mathcal{T}-1)} \sum_{i=1}^{n} \sum_{t=1}^{\mathcal{T}-1} \mathbf{Y}_{*,i,t+1} \mathbf{Y}_{*,i,t}^{\mathrm{T}} \Big] \Big[\frac{1}{n(\mathcal{T}-1)} \sum_{i=1}^{n} \sum_{t=1}^{\mathcal{T}-1} \mathbf{Y}_{*,i,t} \mathbf{Y}_{*,i,t}^{\mathrm{T}} \Big]^{-1}$$

This is a more convenient expression when dealing with high-dimensional data.

References

Abegaz, F. and Wit, E. (2013). Sparse time series chain graphical models for reconstructing genetic networks. *Biostatistics*, **13**(3), 586–599.

Alon, U. (2007). *An Introduction to Systems Biology: Design Principles of Biological Circuits*. Chapman & Hall, London.

Banerjee, O., El Ghaoui, L., and d'Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate Gaussian or binary data. *Journal of Machine Learning Research*, **9**, 485–176.

Bickel, P. J. and Doksum, K. A. (2001). *Mathematical Statistics, Vol. I*. Prentice Hall, Upper Saddle River, New Jersey.

Brockwell, P. J. and Davis, R. A. (2006). *Time Series: Theory and Methods*. Springer, New York.

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007). Pathwise coordinate optimization. *Annals of Applied Statistics*, **1**(2), 302–332.

Fujita, A., Sato, J. R., Garay-Malpartida, H. M., Yamaguchi, R., Miyano, S., Sogayar, M. C., and Ferreira, C. E. (2007). Modeling gene expression regulatory networks with the sparse vector autoregressive model. *BMC Systems Biology*, **1**(39), doi:10.1186/1752–0509–1–39.

Harville, D. A. (2008). *Matrix Algebra From a Statistician's Perspective*. Springer, New York.

Hastie, T. and Tibshirani, R. (2004). Efficient quadratic regularization for expression arrays. *Biostatistics*, **5**(3), 329–340.

Ledoit, O. and Wolf, M. (2004). A well conditioned estimator for largedimensional covariance matrices. *Journal of Multivariate Analysis*, **88**, 365–411.

Lengauer, C., Kinzler, K., and Vogelstein, B. (1998). Genetic instabilities in human cancers. *Nature*, **396**, 623–627.

Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer, Berlin.

Nguyen, D. V., Arpat, A. B., Wang, B., and Carroll, R. J. (2002). Microarray experiments: biological and technical aspects. *Biometrics*, **58**, 701–717.

Nicholls, D. F. and Pope, A. L. (1988). Bias in the estimation of multivariate autoregressions. *Australian Journal of Statistics*, **30A**, 296–309.

Pinkel, D. and Albertson, D. (2005). Array comparative genomic hybridization and its application in cancer. *Nature Genetics*, **37**, S11–S17.

Reale, M. (1998). *A Graphical Modelling Approach to Time Series*. Lancaster University, PhD Thesis, Lancaster, United Kingdom.

Schäfer, J. and Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Application in Genetics and Molecular Biology*, **4**, Article 32.

Vogelstein, B. and Kinzler, K. W. (2004). Cancer genes and the pathways they control. *Nature Medicine*, **10**, 789–799.

Weinberg, R. A. (2006). *The Biology of Cancer*. Garland Science, New York.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. John Wiley & Sons, Chichester, England.

10 Appendix A: Matrix algebra

Theorem 6. Add REF!!!
Suppose a symmetric matrix can be partitioned as:

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{B}^T & \mathbf{D} \end{pmatrix}.$$

Its inverse can than be expressed as:

$$\begin{pmatrix} \mathbf{A}^{-1} + \mathbf{F}\,\mathbf{E}^{-1}\,\mathbf{F}^T & -\mathbf{F}\mathbf{E}^{-1} \\ -\mathbf{E}^{-1}\,\mathbf{F}^T & \mathbf{E}^{-1} \end{pmatrix},$$

where $\mathbf{E} = \mathbf{D} - \mathbf{B}^T\,\mathbf{A}^{-1}\,\mathbf{B}$ and $\mathbf{F} = \mathbf{A}^{-1}\,\mathbf{B}$.

Theorem 7. (*Theorem 13.3.8*, Harville, 2008)
Let $\mathbf{T}$ represent an $m \times m$ matrix, $\mathbf{U}$ an $m \times n$ matrix, $\mathbf{V}$ an $n \times m$ matrix, and $\mathbf{W}$ an $n \times n$ matrix. If $\mathbf{T}$ is nonsingular, then:

$$\begin{vmatrix} \mathbf{T} & \mathbf{U} \\ \mathbf{V} & \mathbf{W} \end{vmatrix} \;=\; |\mathbf{T}|\,|\mathbf{W} - \mathbf{V}\mathbf{T}^{-1}\mathbf{U}|.$$

Theorem 8. (*Theorem 14.2.9*, Harville, 2008)
(Harville, 2008) Let $\mathbf{A}$ represent an $n \times n$ matrix, and $\mathbf{P}$ an $n \times m$ matrix. (1) If $\mathbf{A}$ is nonnegative definite, then $\mathbf{P}^T\,\mathbf{A}\,\mathbf{P}$ is nonnegative definite. (2) If $\mathbf{A}$ is nonnegative definite and rank($\mathbf{P}$) $< m$, then $\mathbf{P}^T\,\mathbf{A}\,\mathbf{P}$ is positive semidefinite. (3) If $\mathbf{A}$ is nonnegative definite and rank($\mathbf{P}$) $= m$, then $\mathbf{P}^T\,\mathbf{A}\,\mathbf{P}$ is positive definite.

Corollary 5. (*Corollary 14.8.6* Harville, 2008)
Suppose that a symmetric matrix $\mathbf{A}$ is partitioned as:

$$\mathbf{A} \;=\; \begin{pmatrix} \mathbf{T} & \mathbf{U} \\ \mathbf{U}^T & \mathbf{W} \end{pmatrix}$$

(where $\mathbf{T}$ and $\mathbf{W}$ are square). Then, $\mathbf{A}$ is positive definite if and only if $\mathbf{T}$ and the Schur complement $\mathbf{W} - \mathbf{U}^T\,\mathbf{T}^{-1}\,\mathbf{U}$ of $\mathbf{T}$ are both positive definite. Similarly, $\mathbf{A}$ is positive definite if and only if $\mathbf{W}$ and the Schur complement $\mathbf{T} - \mathbf{U}\mathbf{W}^{-1}\mathbf{U}^T$ of $\mathbf{W}$ are both positive definite.

Corollary 6. (*Corollary 18.1.7* Harville, 2008)
For any symmetric positive definite matrix $\mathbf{A}$ and for any $n \times n$ symmetric matrix $\mathbf{C}$ such that $\mathbf{C} - \mathbf{A}$ is nonnegative definite,

$$|\mathbf{C}| \;\geq\; |\mathbf{A}|,$$

with equality holding if and only if $\mathbf{C} = \mathbf{A}$.

Theorem 9. (Weyl's monotonicity theorem; **Add ref**)
If $\mathbf{A}$ and $\mathbf{B}$ are $p \times p$, symmetric matrices, and $\mathbf{B}$ is positive semidefinite, then:

$$\lambda_j(\mathbf{A}) \;\leq\; \lambda_j(\mathbf{A} + \mathbf{B}),$$

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

Exercise 33 to Section 14.8 (Harville, 2008) Let $\mathbf{A}$ represent a symmetric nonnegative definite matrix that has been partitioned as

$$\mathbf{A} \quad = \quad \begin{pmatrix} \mathbf{T} & \mathbf{U} \\ \mathbf{V} & \mathbf{W} \end{pmatrix},$$

where $\mathbf{T}$ (and hence $\mathbf{W}$) is square. Show that $\mathbf{V}\,\mathbf{T}^{-}\,\mathbf{U}$ and $\mathbf{U}\,\mathbf{W}^{-}\,\mathbf{V}$ are symmetric and nonnegative definite.

Suppose:
i) $\mathbf{W} - \mathbf{V}\mathbf{T}^{-1}\mathbf{U} \succ 0$ (follows from Corollary 14.8.6).
ii) $\mathbf{V}\mathbf{T}^{-1}\mathbf{U} \succeq 0$
Then, we apply Corollary 18.1.7 with $\mathbf{C} = \mathbf{W}$ and $\mathbf{A} = \mathbf{W} - \mathbf{V}\mathbf{T}^{-1}\mathbf{U}$, and it follows that:

$$|\mathbf{W}| \quad \geq \quad |\mathbf{W} - \mathbf{V}\mathbf{T}^{-1}\mathbf{U}|.$$