Advances in High-Dimensional Covariance Matrix Estimation

vorgelegt von
Diplom-Physiker Daniel Bartz
geboren in Hannover

von der
Fakultät IV – Elektrotechnik und Informatik
der
Technischen Universität Berlin
zur Erlangung des akademischen Grades

Doktor der Naturwissenschaften
– Dr. rer. nat. –
genehmigte Dissertation

Prüfungsausschuss

Vorsitzender  Prof. Dr. Manfred Opper
Gutachter  Prof. Dr. Klaus-Robert Müller
Gutachter  Prof. Dr. Benjamin Blankertz
Gutachter  Prof. Dr. Gilles Blanchard


Berlin, 2016
For Marie Elisa Margarethe
Acknowledgements

This thesis would not have been possible if I had been working alone. I am very grateful for the support I received from my Ph.D. supervisor Prof. Dr. Klaus-Robert Müller. I conducted this research as a member of Klaus’ Machine Learning group at TU Berlin, where he provided a stimulating research environment, encouraged me to follow my research interests and gave me a large degree of freedom.

I am also grateful to Prof. Dr. Peter Bank who supported me during my time at the Quantitative Products Laboratory, allowing me to learn about covariance estimation problems in mathematical finance and econometrics.

I would like to thank Prof. Dr. Benjamin Blankertz and Prof. Dr. Gilles Blanchard who kindly agreed to review the thesis.

Extremely important for this thesis has been the contact to many researchers from various fields. I am very grateful to Steven Lemm, who helped me to get started in science and I would like to thank my Machine Learning for finance collaborators Christian Hesse and Kerr Hatrick. It has been an inspiration to work with Johannes Höhne, whose enthusiasm for Brain-Computer Interfaces led to interesting applications and I profited greatly from mathematical and scientific discussions with Duncan Blythe.

The last years have been made a great experience by all of the present and former members of the Machine Learning group, especially: Sophie Piper, Irene Winkler, Sebastian Bach, Shin-ichi Nakajima, Nikolay Jetchev, Anne Porbadnigk, Tobias Lang, Nico Görnitz, Felix Brockherde, Franz Király, Thorsten Dieckhaus, Mikio Braun, Sven Dähne, Bastian Venthur, Dominik Kühne, Andrea Gerdes, Andreas Ziehe, and the Quantitative Products Lab, Ulrich Horst, Almut-Mirjam Birsner, Marcel Höschler, Antje Fruth, Nicholas Westray, Christine Grün, Markus Mocha, Alexander Groß-Klußman, Peter Grätz, Michael Paulsen, Hartmuth Henkel und Felix Naujokat.

During the last few years, the support of my family and friends, especially that of my parents, my brother and my wife, has been extremely important to me.
Abstract

Many applications require precise estimates of high-dimensional covariance matrices. The standard estimator is the sample covariance matrix, which is conceptually simple, fast to compute and has favorable properties in the limit of infinitely many observations. The picture changes when the dimensionality is of the same order as the number of observations. In such cases, the eigenvalues of the sample covariance matrix are highly biased, the condition number becomes large and the inversion of the matrix gets numerically unstable.

A number of alternative estimators are superior in the high-dimensional setting, which include as subcategories structured estimators, regularized estimators and spectrum correction methods. In this thesis I contribute to all three areas.

In the area of structured estimation, I focus on models with low intrinsic dimensionality. I analyze the bias in Factor Analysis, the state-of-the-art factor model and propose Directional Variance Adjustment (DVA) Factor Analysis, which reduces bias and yields improved estimates of the covariance matrix. Analytical shrinkage of Ledoit and Wolf (LW-Shrinkage) is the most popular regularized estimator. I contribute in three aspects: first, I provide a theoretical analysis of the behavior of LW-Shrinkage in the presence of pronounced eigendirections, a case of great practical relevance. I show that LW-Shrinkage does not perform well in this setting and propose aoc-Shrinkage which yields significant improvements. Second, I discuss the effect of autocorrelation on LW-Shrinkage and review the Sancetta-Estimator, an extension of LW-Shrinkage to autocorrelated data. I show that the Sancetta-Estimator is biased and propose a theoretically and empirically superior estimator with reduced bias. Third, I propose an extension of shrinkage to multiple shrinkage targets. Multi-Target Shrinkage is not restricted to covariance estimation and allows for many interesting applications which go beyond regularization, including transfer learning. I provide a detailed theoretical and empirical analysis.

Spectrum correction approaches the problem of covariance estimation by improving the estimates of the eigenvalues of the sample covariance matrix. I discuss the state-of-the-art approach, Nonlinear Shrinkage, and propose a cross-validation based covariance (CVC) estimator which yields competitive performance at increased numerical stability and greatly reduced complexity and computational cost. On all data sets considered, CVC is on par or superior in comparison to the regularized and structured estimators.

In the last chapter, I conclude with a discussion of the advantages and disadvantages of all covariance estimators presented in this thesis and give situation-specific recommendations. In addition, the appendix contains a systematic analysis of Linear Discriminant Analysis as a model application, which sheds light on the interdependency between the generative model of the data and various covariance estimators.
Zusammenfassung

Viele Anwendungen benötigen Schätzungen hochdimensionaler Kovarianzmatrizen. Der Standardschätzer ist die Stichprobenkovarianz. Die Stichprobenkovarianz ist konzeptionell einfach, schnell zu berechnen und hat gute Eigenschaften im Grenzwert unendlich vieler Beobachtungen. Dies ändert sich, wenn die Dimensionalität die selbe Größenordnung wie die Anzahl der Beobachtungen hat. Dann weisen die Eigenwerte der Stichprobenkovarianz einen hohen systematischen Fehler auf, die Kondition wird groß und das Inverse numerisch instabil.


Im Bereich der **strukturierten Schätzer** konzentriere ich mich auf Modelle mit geringer intrinsischer Dimensionalität. Ich analysiere den systematischen Fehler der Faktoranalyse, dem Stand der Technik im Bereich Faktormodelle, und schlage **Directional Variance Adjustment (DVA)**-Faktoranalyse vor, welche den systematischen Fehler korrigiert und eine verbesserte Kovarianzschatzung liefert.


# Contents

Acknowledgements ................................................................. v
Abstract .................................................................................. vii
Zusammenfassung ................................................................. ix
Contents ................................................................................... xi
List of Figures ........................................................................ xv
List of Tables ........................................................................... xvii
Symbols and Notation .............................................................. xix

## 1 Introduction
1.1 From epistemology to correlations ...................................... 1
1.2 From bivariate correlations to covariance matrices .......... 1
1.3 Where are covariance matrices needed? ......................... 3
1.4 Covariance matrix estimation ........................................... 4
1.4.1 The sample covariance matrix ............................... 4
1.4.2 On the difficulty of covariance estimation ................. 4
1.4.3 Asymptotic settings ............................................ 5
1.4.4 Evaluation of covariance estimators .................. 6
1.5 Approaches to covariance estimation ............................ 7
1.6 Application domains contained in this thesis .................. 8
1.6.1 Stock market data ............................................ 8
1.6.2 EEG data ................................................... 8
1.6.3 Audio data: spoken letters ................................ 10
1.6.4 Image data: handwritten digits .......................... 10
1.7 Contributions .............................................................. 10

## 2 Structured Estimation
2.1 Factor models as structured covariance estimators .......... 13
2.1.1 Principal Component Analysis (PCA) .................... 15
2.1.2 Probabilistic PCA and Factor Analysis ................. 15
2.1.3 Model selection for the number of factors ............ 17
2.2 Directional Variance Adjustment .................................. 17
2.2.1 Systematic error in Factor Analysis .................... 17
2.2.2 Directional Variance Adjustment: correcting the systematic error 18
2.2.3 Simulation results ........................................... 20
2.3 Evaluation in a portfolio simulation ........................................... 21
  2.3.1 Approaches compared in the simulations ................................. 22
  2.3.2 Design of portfolio simulations ........................................... 24
  2.3.3 Portfolio simulation results .............................................. 26
  2.3.4 Conclusions from portfolio simulations ................................. 29
2.4 Chapter summary ................................................................. 30

3 Regularized Estimation ......................................................... 31
  3.1 Introduction to analytic shrinkage ......................................... 32
  3.2 Shrinkage for arbitrary covariance structures ........................... 35
    3.2.1 Implicit assumptions on the covariance structure .................. 35
    3.2.2 LW-Shrinkage for arbitrary covariance structures ................ 37
    3.2.3 LW-Shrinkage in the LDL .............................................. 38
    3.2.4 LW-Shrinkage in the FOLDL ......................................... 39
    3.2.5 Structural adaptation: automatic orthogonal complement Shrinkage .... 40
    3.2.6 Simulation results .................................................... 45
    3.2.7 Real world data results ............................................. 48
  3.3 Covariance shrinkage for autocorrelated data ............................ 51
    3.3.1 Extension of LW-Shrinkage for autocorrelated data ................. 51
    3.3.2 Theoretical results ................................................ 53
    3.3.3 Simulations ......................................................... 55
    3.3.4 Real world data: BCI based on motor imagery ...................... 56
3.4 Chapter summary ................................................................. 58

4 Regularized Estimation II: MTS ............................................... 61
  4.1 Introduction to Multi-Target Shrinkage .................................... 61
  4.2 MTS notation and distributional assumptions ............................. 64
  4.3 General Multi-Target Shrinkage ............................................ 65
    4.3.1 Estimation of Multi-Target Shrinkage ................................ 66
    4.3.2 Consistency of MTS of the mean ................................... 68
    4.3.1 Consistency of MTS of the mean ................................... 68
  4.4 Multi-Target Shrinkage of the mean ...................................... 68
    4.4.1 Consistency of MTS of the mean ................................... 68
  4.5 Multi-Target Shrinkage of the covariance matrix ........................ 69
    4.5.1 Consistency of MTS of the covariance ................................ 70
  4.6 Simulations ........................................................................ 71
    4.6.1 Simulations for MTS of the mean ................................... 71
    4.6.2 Simulations for MTS of the covariance ................................ 74
  4.7 Multi-Target Shrinkage on real world data ................................ 77
    4.7.1 MTS of the mean for ERP-BCI ....................................... 77
    4.7.2 MTS of the covariance matrix for MI-BCI ............................ 77
  4.8 Chapter Summary ................................................................. 78

5 Spectrum Correction ............................................................. 81
  5.1 Improved estimation by spectrum correction .............................. 82
  5.2 Random Matrix Theory (RMT) for spectrum correction ................. 82
    5.2.1 Inverting the Marčenko-Pastur law .................................. 83
    5.2.2 Adapting RMT to the covariance estimation problem .............. 84
  5.3 Using cross-validation for covariance estimation ....................... 85
    5.3.1 Precision matrix estimation .......................................... 85
    5.3.2 Estimation error of the CVC approach ................................ 86
    5.3.3 Runtime comparison .................................................. 89
  5.4 Real world data .................................................................. 90
5.5 Chapter summary ................................................................. 91

6 Summary and Outlook ......................................................... 93
  6.1 Open questions and future work ......................................... 94
  6.2 Which estimator to use in practice ...................................... 95

A Estimation Error vs. Performance ........................................ 99
  A.1 Discriminative information and covariance ............................. 101
  A.2 Simulation design ............................................................ 103
  A.3 Simulation results ........................................................... 104
  A.4 Appendix summary .......................................................... 108

B Proofs and Derivations for Shrinkage and Structure ................. 109
  B.1 Proof of Theorem 3.1 (Largest eigenvalue growth rate) ............. 109
  B.2 Proof of Theorem 3.2 (Dispersion growth rate) ....................... 109
  B.3 Proof of Theorem 3.3 (Sparsity in the limit) ......................... 110
  B.4 Proof of Theorem 3.4 (LDL behavior of LW-Shrinkage) ............ 110
  B.5 Proof of Theorem 3.5 (LDL consistency of LW-Shrinkage) .......... 112
  B.6 Proof of Theorem 3.6 (FOLDL behavior of LW-Shrinkage) ......... 116
  B.7 Proof of Theorem 3.7 (FOLDL consistency of LW-Shrinkage) ....... 117
  B.8 Proof of Theorem 3.8 (Convergence of the eigendecomposition) .. 119
  B.9 Proof of Theorem 3.9 (Consistency of oc-shrinkage) ............... 121
  B.10 Proof of Theorem 3.10 (Convergence of the error) ................. 123
    B.10.1 Additional lemmas .................................................... 124
  B.11 Conservative estimate of the oc-shrinkage improvement ............ 134
    B.11.1 Variance of $\hat{\nu}_{\text{oc}}^2 = \frac{1}{p} \sum_{i,j} \left\{ \text{Var} \left( \hat{S}_{ij}^2 \right) - \text{Cov} \left( \hat{S}_{ij}^2, \hat{\nu}_{\text{oc}} I_{ij} \right) \right\}$ .......... 134
    B.11.2 Variance of $\hat{E}_{\text{oc}}^2 = \frac{1}{p} \sum_{i,j} (\hat{S}_{ij}^2 - \hat{\nu}_{\text{oc}} I_{ij})^2$ ............... 136
    B.11.3 Covariance between $\hat{V}_{\text{oc}}$ and $\hat{E}_{\text{oc}}$ .............. 137
    B.11.4 Calculation of the appearing covariance expressions .......... 138

C Proofs for Shrinkage and autocorrelation ............................ 143
  C.1 Proof of Theorem 3.11 (FOLDL consistency under AC) ............... 143

D Proofs for Multi-Target Shrinkage ....................................... 147
  D.1 Proof of Theorem 4.1 (MTS quadratic program) ....................... 147
  D.2 Proof of Theorem 4.2 (Consistency of MTS) .......................... 147
  D.3 Proof of Theorem 4.3 (LDL consistency of MTS of the mean) ....... 149
  D.4 Proof of Theorem 4.4 (FOLDL consistency of MTS of the mean) .... 152
  D.5 Proof of Theorem 4.5 (LDL consistency of MTS of the covariance) .. 153
  D.6 Proof of Theorem 4.6 (FOLDL consistency of MTS of the covariance) 158

Bibliography ................................................................. 161
# List of Figures

1.1 First scatter plot from Galton ........................................... 2
1.2 Systematic error in the spectrum of the sample covariance matrix ........................................... 5

2.1 Two-dimensional example of a 1-factor model ........................................... 16
2.2 Systematic error of Factor Analysis ........................................... 19
2.3 Illustration of the DVA algorithm ........................................... 21
2.4 Systematic error in standard and DVA Factor Analysis ........................................... 22
2.5 Absolute relative error for standard and DVA Factor Analysis ........................................... 22
2.6 Absolute relative error for standard and DVA Factor Analysis in high dimensions ........................................... 23
2.7 Regularization dependency of portfolio risk in the US market ........................................... 27
2.8 Regularization dependency of portfolio risk in the EU market ........................................... 28
2.9 Regularization dependency of portfolio risk in the HK market ........................................... 28

3.1 Simulation on eigenvalue dispersion and dimensionality ........................................... 37
3.2 Empirical eigenvalue dispersion dependency on dimensionality ........................................... 37
3.3 Illustration of orthogonal complement Shrinkage ........................................... 42
3.4 PRIAL vs. dimensionality in Simulation 1 ........................................... 46
3.5 PRIAL vs. correlation probability in Simulation 2 ........................................... 46
3.6 PRIAL vs. correlation probability on the orthogonal complement in Simulation 2 ........................................... 47
3.7 Condition number in Simulation 2 ........................................... 47
3.8 PRIAL vs. dimensionality in Simulation 3 ........................................... 48
3.9 High-variance components responsible for failure of LW-Shrinkage ........................................... 50
3.10 Dependency of the eigendecomposition on autocorrelation ........................................... 51
3.11 Dependency of shrinkage on autocorrelation ........................................... 52
3.12 Dependence of the variance estimates on the dimensionality ........................................... 55
3.13 Robustness to the choice of lag parameter $b$ ........................................... 56
3.14 BCI motor imagery data for lag parameter $b = 75$ and $b = 300$ ........................................... 57
3.15 Subject-wise BCI classification accuracies and time demands ........................................... 57

4.1 Geometric illustration of Multi-Target Shrinkage ........................................... 62
4.2 Geometric illustration of Multi-Target Shrinkage for handwritten digits ........................................... 63
4.3 Decay of the squared error for increasing number of shrinkage targets ........................................... 63
4.4 Large-dimensional limit (LDL) of MTS of the mean to additional data sets ........................................... 72
4.5 Finite observations large-dimensional limit (FOLDL) of MTS of the mean to additional data sets ........................................... 73
4.6 Accuracy gain for MTS for Linear Discriminant Analysis ........................................... 74
4.7 Large-dimensional limit (LDL) of MTS of the covariance to additional data sets ........................................... 74
4.8 Finite observations large-dimensional limit (FOLDL) of MTS of the covariance to additional data sets ........................................... 75
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.9</td>
<td>MTS of the covariance to identity and additional data sets</td>
<td>75</td>
</tr>
<tr>
<td>4.10</td>
<td>Accuracy gain for MTS of the covariance for CSP</td>
<td>76</td>
</tr>
<tr>
<td>4.11</td>
<td>Classification accuracy of the ERP-BCI data using several estimates of the mean</td>
<td>77</td>
</tr>
<tr>
<td>4.12</td>
<td>Dependency on the number of training trials of motor imagery BCI</td>
<td>78</td>
</tr>
<tr>
<td>4.13</td>
<td>Subject-wise classification accuracies for motor imagery BCI</td>
<td>78</td>
</tr>
<tr>
<td>5.1</td>
<td>Illustration of the estimation error of loo-CVC</td>
<td>87</td>
</tr>
<tr>
<td>5.2</td>
<td>Estimation error of CVC estimators</td>
<td>89</td>
</tr>
<tr>
<td>5.3</td>
<td>Runtime comparison of NLS and CVC</td>
<td>89</td>
</tr>
<tr>
<td>A.1</td>
<td>Shrinkage failure in the presence of a noise channel</td>
<td>100</td>
</tr>
<tr>
<td>A.2</td>
<td>2D illustration: average predictions</td>
<td>101</td>
</tr>
<tr>
<td>A.3</td>
<td>2D illustration: single realization</td>
<td>102</td>
</tr>
<tr>
<td>A.4</td>
<td>SNR and eigenvalues in the main simulation and variation 1</td>
<td>103</td>
</tr>
<tr>
<td>A.5</td>
<td>SNR and eigenvalues in variation 2</td>
<td>104</td>
</tr>
<tr>
<td>A.6</td>
<td>Main simulation: systematic comparison of classification accuracies</td>
<td>105</td>
</tr>
<tr>
<td>A.7</td>
<td>Comparison of (spectrum) oracle and precision oracle</td>
<td>106</td>
</tr>
<tr>
<td>A.8</td>
<td>Variation 1: systematic comparison for low Bayes-optimal accuracy</td>
<td>106</td>
</tr>
<tr>
<td>A.9</td>
<td>Variation 2: systematic comparison for non-monotonic SNR</td>
<td>107</td>
</tr>
</tbody>
</table>
### List of Tables

1. Overview of the notation .......................................................... xx
2.1 DVA results: portfolio risk ...................................................... 27
2.2 DVA results: portfolio risk under regularization ......................... 29
3.1 Results on consistency and limit behavior ............................... 41
3.2 aoc-Shrinkage results: portfolio risk ..................................... 48
3.3 aoc-Shrinkage accuracies for classification tasks on ISOLET and USPS data ......................................................... 49
3.4 Accuracies for classification tasks on NU-ERP-BCI data .......... 50
4.1 General, mean and covariance MTS ....................................... 65
5.1 CVC results: portfolio risk ..................................................... 90
5.2 CVC, CV-Shrinkage and FM accuracies for ISOLET and USPS data ............................................................... 91
5.3 CVC, CV-Shrinkage and FM accuracies for the NU-ERP-BCI data ............................................................. 91
6.1 Covariance matrix estimators ............................................... 96
Symbols and Notation

The notation adheres to the following conventions:

- Matrices $\mathbf{M}$ and vectors $\mathbf{v}$ are written in upper case and lower case bold letters, respectively; their entries are given by $M_{ij}$ and $v_i$. $\mathbf{m}_j$ denotes the $j^{th}$ column of the matrix $\mathbf{M}$ with entries $m_{ij} \equiv M_{ij}$.
- $\mathbf{M}^\top$ denotes the transpose of a matrix $\mathbf{M}$.
- Quantities with a hat, $\hat{\mathbf{M}}$ and $\hat{\mathbf{v}}$, always denote estimators.
- $\text{Var}(a)$ and $\text{Cov}(a, b)$ denote the variance of $a$ and the covariance between $a$ and $b$, respectively.
- For asymptotic behavior, the Bachmann-Landau symbols $O$, $o$ and $\Theta$ are used. These are defined in the following way:
  (i) $O$ denotes asymptotically bounded from above:
      
      \[
      f = O(g) \iff \exists C > 0 \exists x_0 > 0 \forall x > x_0 : |f(x)| \leq C \cdot |g(x)|.
      \]

  (ii) $o$ denotes asymptotically dominated:

      \[
      f = o(g) \iff \forall C > 0 \exists x_0 > 0 \forall x > x_0 : |f(x)| \leq C \cdot |g(x)|.
      \]

  (iii) $\Theta$ denotes asymptotically bounded from above and below:

      \[
      f = \Theta(g) \iff \exists c > 0 \exists C > 0 \exists x_0 > 0 \forall x > x_0 : c \cdot |g(x)| \leq |f(x)| \leq C \cdot |g(x)|.
      \]

A complete list of notation for this thesis is summarized on the next page in Table 1.
### Table 1. Overview of the notation.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>$V_{X,Y}$</td>
<td>covariance between two random variables $X$ and $Y$</td>
</tr>
<tr>
<td>$\rho_{X,Y}$</td>
<td>correlation between two random variables $X$ and $Y$</td>
</tr>
<tr>
<td>$\sigma_X$</td>
<td>standard deviation of a variable $X$</td>
</tr>
<tr>
<td>$E[X]$</td>
<td>expected value of a random variable $X$</td>
</tr>
<tr>
<td>$n$</td>
<td>number of observations</td>
</tr>
<tr>
<td>$X$, $x_{it}$</td>
<td>observations, $(p \times n)$-matrix</td>
</tr>
<tr>
<td>$\mu$, $\overline{\mu}$</td>
<td>population mean and sample mean</td>
</tr>
<tr>
<td>symbol</td>
<td>estimate of quantity $\text{symbol}$ calculated on the data</td>
</tr>
<tr>
<td>$p$</td>
<td>dimensionality / index of the sequence of models</td>
</tr>
<tr>
<td>$X$, $x_{it}$</td>
<td>observations, $(p \times n)$-matrix</td>
</tr>
<tr>
<td>$\mathbf{C}$, $\mathbf{S}$</td>
<td>population covariance and sample covariance</td>
</tr>
<tr>
<td>$d_f$</td>
<td>degrees of freedom in $\mathbf{C}$</td>
</tr>
<tr>
<td>$\mathbf{v}_1, \ldots, \mathbf{v}_p$</td>
<td>eigenvectors of $\mathbf{C}$</td>
</tr>
<tr>
<td>$\mathbf{v}_1, \ldots, \mathbf{v}_p$</td>
<td>eigenvectors of $\mathbf{S}$</td>
</tr>
<tr>
<td>$\gamma_1, \ldots, \gamma_p$</td>
<td>eigenvalues of $\mathbf{C}$</td>
</tr>
<tr>
<td>$\tilde{\gamma}_1, \ldots, \tilde{\gamma}_p$</td>
<td>eigenvalues of $\mathbf{S}$</td>
</tr>
<tr>
<td>$c$</td>
<td>concentration $n/p$ (ratio between number of observations and dimensionality)</td>
</tr>
<tr>
<td>$f = O(g)$</td>
<td>$f$ asymptotically bounded from above and below by $g$</td>
</tr>
<tr>
<td>$f = o(g)$</td>
<td>$f$ asymptotically dominated by $g$</td>
</tr>
</tbody>
</table>

### Chapter 1 Introduction and Motivation
- $X \perp Y$: random variables $X$ and $Y$ are independent
- $a \perp b$: the vectors $a$ and $b$ are orthogonal
- $M$: number of factors in a factor model
- $f$: Factors, $M$-dimensional vector
- $\mathbf{f}$: noise, $p$-dimensional vector
- $\mathbf{W}$: mixing matrix, $(p \times M)$-matrix
- $\eta_{i,\text{FA}}$: systematic relative error of Factor Analysis
- $\xi_{i,\text{FA}}$: systematic absolute relative error of Factor Analysis
- $\mathcal{F}$: estimated Factor Analysis parameter set ($\hat{\mathbf{W}}$ and $\hat{\mathbf{C}}$)

### Chapter 2 Structured Estimation
- $\mathbf{Y}$, $y_{it}$: observations in the population eigenbasis ($p \times n$ matrix)
- $\mathbf{S}$: sample covariance in the population eigenbasis
- $\text{Tr}(\mathbf{M})$: trace of the matrix $\mathbf{M}$
- $\mathbf{R}$: rotation into the eigenbasis ($p \times p$ matrix)
- $\Delta$: expected squared error (ESE) of the shrinkage estimator
- $\hat{\Delta}$: squared error (SE) of the shrinkage estimator
- $\lambda$: shrinkage intensity
- $Q_p$: set of all quadruples consisting of distinct integers between 1 and $p$
- $|Q_p|$: cardinality of $Q_p$
- $d_n$: eigenvalue dispersion
- $k$: parameterization of the normalized eigenvalue dispersion
- $\eta_i$: normalized observations in the eigenbasis
- $\delta_{\hat{\mathbf{C}}}$: normalized improvement in ESE on the orthogonal complement
- $\delta_{\hat{\mathbf{C}}}$: normalized improvement in SE on the orthogonal complement
\( \Upsilon_{ij}(s) \) \hspace{1cm} \text{autocovariances of second order terms} \( (n-s)/n \text{Cov}(x_{it}x_{jt}, x_{i,t+s}x_{j,t+s}) \)

\( b \) \hspace{1cm} \text{number of lags included in AC-consistent shrinkage}

<table>
<thead>
<tr>
<th>Chapter 4</th>
<th>Regularized Estimation II: MTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>( q )</td>
<td>number of parameters</td>
</tr>
<tr>
<td>( \theta )</td>
<td>set of parameters</td>
</tr>
<tr>
<td>( \hat{\theta} )</td>
<td>unbiased estimate of the set of parameters</td>
</tr>
<tr>
<td>( \tau_{\hat{\theta}} )</td>
<td>limit behavior of the unbiased estimator (G1)</td>
</tr>
<tr>
<td>( \Delta \hat{\theta} )</td>
<td>expected squared error, here of the unbiased estimator</td>
</tr>
<tr>
<td>( \tau_\gamma )</td>
<td>limit behavior of the average squared eigenvalue (D2)</td>
</tr>
<tr>
<td>( Z )</td>
<td>observations in the eigenbasis of a different data set (( p \times n ) matrix)</td>
</tr>
<tr>
<td>( (\text{symbol})^k )</td>
<td>quantity symbol on the data set ( k )</td>
</tr>
<tr>
<td>( \alpha_4, \beta_4 )</td>
<td>bounds on the ratio between second and fourth moments (D3)</td>
</tr>
<tr>
<td>( \alpha_8, \beta_8 )</td>
<td>bounds on the ratio between fourth and eighth moments (D4)</td>
</tr>
<tr>
<td>( K )</td>
<td>number of shrinkage targets</td>
</tr>
<tr>
<td>( \lambda_k )</td>
<td>shrinkage intensity of the ( k^{th} ) shrinkage target</td>
</tr>
<tr>
<td>( A )</td>
<td>matrix containing estimates of the quality of the targets</td>
</tr>
<tr>
<td>( b )</td>
<td>vector containing variance of sample estimate and correlation to targets</td>
</tr>
<tr>
<td>( \tau_k^A )</td>
<td>limit behavior containing variance of sample estimate and correlation to targets</td>
</tr>
<tr>
<td>( \tau_k^A )</td>
<td>limit behavior of the quality of target ( k ) (G3)</td>
</tr>
<tr>
<td>( \tau_k^C )</td>
<td>limit behavior of the quality of the mean of data set ( k ) (M1)</td>
</tr>
<tr>
<td>( \tau_k^C )</td>
<td>limit behavior of the quality of the covariance of data set ( k ) (C1)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter 5</th>
<th>Spectrum Correction</th>
</tr>
</thead>
<tbody>
<tr>
<td>( H_p, H_\infty )</td>
<td>distribution of population eigenvalues at dim. ( p ) / in the limit</td>
</tr>
<tr>
<td>( F_p, F_\infty )</td>
<td>distribution of sample eigenvalues at dim. ( p ) / in the limit</td>
</tr>
<tr>
<td>( G_p, G_\infty )</td>
<td>distribution of eigenvalues of ( n^{-1}X^\top X ) at dim. ( p ) / in the limit</td>
</tr>
<tr>
<td>( m_{H/F/G} )</td>
<td>Stieltjes transform of ( H, F ) or ( G )</td>
</tr>
<tr>
<td>( \gamma^* )</td>
<td>(spectrum) oracle</td>
</tr>
<tr>
<td>( \gamma^0 )</td>
<td>precision oracle</td>
</tr>
<tr>
<td>( \gamma^\text{NLS} )</td>
<td>(spectrum) oracle Nonlinear Shrinkage estimator</td>
</tr>
<tr>
<td>( \gamma^\text{NLS} )</td>
<td>precision oracle Nonlinear Shrinkage estimator</td>
</tr>
<tr>
<td>( \gamma^{\text{NLS-precision}} )</td>
<td>precision oracle Nonlinear Shrinkage estimator</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 From epistemology to correlations

Epistemology deals with the nature of knowledge: what is knowledge and how can it be acquired? The natural sciences are closely related to empiricism, claiming that knowledge comes from sensory observations. For an empiricist, the mere existence of objects is easy: I see an object, hence, there is an object.

The mechanics of the world cannot be observed and are therefore a more complex issue. The only thing we can observe are co-occurrences which allow us to infer dependencies, construct causal theories and make predictions about the future. For a neolithic human, the knowledge that sudden silence indicates a nearby predator was quite helpful.

The quantification of co-occurrences has a long history (Pearson, 1920; Lee Rodgers & Nicewander, 1988). The first roots for this can be seen in Gauss’ 1823 work on the normal surface of $N$ correlated variates and the related work by the French astronomer Auguste Bravais in 1868, who studied bivariate normal distributions. Many scientists investigated co-occurrences. In 1843, the British philosopher John Stuart Mill discussed the relation between induction and concomitant variation, and by 1868 Charles Darwin frequently used the term correlation to describe connections between parts of an organism.

A big step forward was taken by Sir Francis Galton (1877) who introduced reversion. Reversion, and the later successor regression, are both tightly related to correlation. Galton (1886) also published the first bivariate scatter plot (see Figure 1.1). Later, Galton (1889) defined the correlation coefficient $r$ conceptually. Galton’s work was based entirely on quantiles. His work was taken up by the statistician Karl Pearson, who formally defined the (Pearson) correlation coefficient in 1895 as the standardized covariance (Pearson, 1895):

\[
\text{covariance: } \quad V_{X,Y} := \mathbb{E}[(X - \mu_X)(Y - \mu_Y)] \\
\text{correlation: } \quad \rho_{X,Y} := \frac{V_{X,Y}}{\sigma_X \sigma_Y} \tag{1.2}
\]

Here, $X$ and $Y$ are random variables, $\mathbb{E}$ denotes the expectation, $\mu_Z := \mathbb{E}[Z]$ and $\sigma_Z := \sqrt{\mathbb{E}[(Z - \mu_Z)^2]}$ denote the population mean and the standard deviation of a random variable $Z$, respectively.

1.2 From bivariate correlations to covariance matrices

Many applications in the modern world feature more than two variables. The generalization of eq. (1.2) to a set of $p$ variables $X_1, X_2, \ldots, X_p$ is straightforward: the correlation matrix $\rho$ is defined by a $(p \times p)$-matrix whose entry $i, j$ is given by the bivariate correlation $\rho_{X_i,X_j}$. 
In this thesis, I focus on the covariance matrix $C$, which is of more interest in many applications (see next section). It is defined by:

$$ C_{ij} := \mathbb{E} \left[ (X_i - \mu_{X_i})(X_j - \mu_{X_j}) \right]. $$

The extension from an univariate to a matrix-valued statistic is straightforward, yet the matrix is a far more interesting and complex object. There are a number of characteristics which come into play when moving from coefficient to matrix:

- **Degrees of freedom:** the number of parameters increases quadratically in $p$. The covariance matrix has

  $$\text{df}_C = \frac{1}{2}p(p + 1)$$

  degrees of freedom.

- **Positive semi-definiteness:** the set of covariance matrices is a specific subset of all $(p \times p)$-matrices, the set of positive semi-definite matrices:

  $$ C \in \mathbb{R}^{p \times p} : a^T C a \geq 0, \forall a \in \mathbb{R}^p. $$

- **Eigenvalues and eigenvectors:** the eigenvalues $\gamma_i$ and eigenvectors $v_i$ play an important role in many applications. They are defined by the equation

  $$ C v_i = \gamma_i v_i, \quad s.t. \|v_i\|^2 = 1. $$

  This directly yields $\gamma_i = v_i^T C v_i$, which implies that the eigenvalues of a positive semi-definite matrix are greater than or equal to zero.

- **Rank and invertibility:** in many applications, the inverse of the covariance matrix is important. If the (estimate of the) covariance matrix does not have full rank, it is not invertible.

- **Condition number:** even if a matrix has full rank, numerical inversion may be unstable: if the condition number (the ratio between largest and smallest eigenvalue) is large, small estimation errors in the covariance may result in large estimation errors in the inverse.

These aspects play a major role in covariance matrix estimation and are discussed throughout the thesis.
1.3 WHERE ARE COVARIANCE MATRICES NEEDED?

Covariance matrices play an important role in many machine learning algorithms and estimation procedures in statistics, which in turn have many applications. Here is a list of some of the most prominent cases:

- **Linear Discriminant Analysis (LDA)** is a classification algorithm based on modeling class distributions as multivariate normals\(^1\) (Fisher, 1936; Friedman, 1989; Hastie et al., 2008), which requires the estimation of covariance matrices. LDA assumes that the covariance matrices of all classes are identical.

- **Quadratic Discriminant Analysis (QDA)** only differs in one respect from LDA: the restriction of equal class covariance matrices is removed and one covariance matrix per class is estimated (Hastie et al., 2008). Hence, for the estimation of each covariance matrix, less data points are available.

- **Common Spatial Patterns (CSP)** is a preprocessing method popular in Brain-Computer Interfaces based on motor imagery (Lotte et al., 2007; Blankertz et al., 2008). CSP searches for spatial directions which maximize the ratio of variances between two experimental conditions.

- **Portfolio optimization** in stock markets (Markowitz, 1952) and **risk measurement** (Jorion, 1996) are applications in quantitative finance which rely on covariance estimates. The dependency between assets is captured in the covariance.

  Similar methods are highly relevant for the insurance industry (Cummins, 2000) and energy markets (Denton et al., 2003).

- **Gene relevance networks** (Butte et al., 2000) and **gene association networks** (Schäfer & Strimmer, 2005) are graphical models of the dependence structure of genes. Relevance networks are directly based on pairwise correlations, association networks are inferred from the covariance matrix.

- **Adaptive beamforming** deals with the extraction of a signal of interest measured by an array of sensors. The **Capon Beamformer** linearly combines the sensors to minimize the noise in the signal (Capon, 1969; Abrahamsson et al., 2007). Mathematically, this is closely related to portfolio optimization.

  Modern wireless communication protocols like **multi-carrier code-division multiple access (MC-CDMA)** optimize bandwidth based on covariance estimates (Tulino & Verdú, 2004).

- **Policy learning** in the field of robotics often requires estimates of the covariance matrix between policy parameters (Deisenroth et al., 2013).

- **Principal Component Analysis (PCA)** (Jolliffe, 2002) and its extensions (Schölkopf et al., 1998; Zou et al., 2006) are popular techniques for dimension reduction and denoising.

- **Independent Component Analysis (ICA)** relies on pre-whitening based on the covariance matrix (Hyvärinen et al., 2002).

- The **Generalized Method of Moments (GMM)** is a method for estimating the parameters in a statistical model, where prior knowledge and constraints are formulated as moment conditions (Hansen, 1982). Precise covariance estimates increase the efficiency of the GMM.

- **Generalized Least Squares (GLS)** regression differs from ordinary least squares (OLS) regression by allowing for dependent noise. Hence it requires the estimation of the noise covariance matrix (Kariya & Kurata, 2004).

---

\(^1\)In the alternative derivation by Fisher, the separation maximizes the ratio of between-class-variance to within-class-variance.
CHAPTER 1. INTRODUCTION

- Canonical Correlation Analysis (CCA) finds maximally correlated linear combinations of two vector-valued random variables (Hotelling, 1936). The estimation relies on covariances and cross-covariances.

1.4 Covariance matrix estimation

1.4.1 The sample covariance matrix

The standard estimator of the population covariance is the sample covariance matrix $\hat{S}$. It is obtained by replacing expectations in eq. (1.3) by sample estimates:

$$\hat{S}_{ij} := \frac{1}{n-1} \sum_{t=1}^{n} (x_{it} - \hat{\mu}_i) \cdot (x_{jt} - \hat{\mu}_j).$$  (1.4)

Here $X$ is the $(p \times n)$-matrix containing $n$ observations of $p$ variables and

$$\hat{\mu}_i := \frac{1}{n} \sum_{t=1}^{n} x_{it}$$

is the sample mean. The sample covariance matrix, has appealing properties: it is easy to calculate and its entries are unbiased and consistent estimates of the entries in the population covariance matrix: for $n \to \infty$, the sample covariance matrix converges to the population covariance matrix (Hastie et al., 2008). These properties make it popular and widely used.

1.4.2 On the difficulty of covariance estimation

For large $n$, the errors in the entries of the sample covariance matrix are small. Unfortunately, the properties of the matrix also depend on the dimensionality $p$. In high-dimensional as well as small sample size settings the performance of the sample covariance matrix degrades. As these settings lack a clear definition and are interdependent, the difficulty of the task of covariance estimation is commonly characterized by the concentration $c := n/p$, the ratio of sample size to dimensionality. Small values of $c$ indicate high dimensionality or small sample sizes.

A central result from Random Matrix Theory (RMT) shows that the spectrum of a random matrix is governed by the concentration $c$ (Edelman & Rao, 2005; Bai & Silverstein, 2010). To be more precise, let us assume that the population eigenvalues are drawn from a fixed distribution. Then, for both $n$ and $p$ large, the distribution of the sample eigenvalues only depends on the concentration $c$. The density of sample eigenvalues in the large-dimensional limit, $n, p \to \infty$ and $n/p \to c$, is given by the Marchenko-Pastur Law (Marchenko & Pastur, 1967). Already for sample sizes as small as 20 or 30, the distribution of sample eigenvalues is well-approximated by the distribution in the limit (Tulino & Verdú, 2004).

Figure 1.2 shows the spectrum of the sample covariance matrix for various ratios of sample sizes to dimensionality and all population eigenvalues equal to 1. It can be seen that there is a systematic error: Large eigenvalues are over- and small eigenvalues are underestimated (see, e.g. Friedman (1989)). The smaller the ratio $c$, the larger the magnitude of the systematic error. For the degenerate case ($c < 1$), $p - n$ eigenvalues are zero. Even for $c = 100$, the empirical spectrum differs visibly from the population spectrum.

Many applications requiring covariance matrices rely on the inverse of the covariance matrix, also called the precision matrix. For ratios $c$ not much larger than 1, the sample covariance becomes ill-conditioned, implying that its inverse incurs large estimation errors. When the number of observations falls below the number of dimensions, $c < 1$, the covariance matrix becomes singular.

\[2\] If normalized by $n$ instead of $n-1$, the sample covariance is maximum likelihood under normality but no longer unbiased.
1.4. COVARIANCE MATRIX ESTIMATION

Figure 1.2. Systematic error in the spectrum of the sample covariance matrix for different ratios of sample size to dimensionality.

1.4.3 Asymptotic settings

The standard asymptotic setting, where \( p \) is fixed while \( n \) tends to infinity, is of little interest for research on covariance estimation. In this setting, the sample covariance matrix is a consistent estimator. Hence, settings are analyzed in which the dimensionality \( p \) also tends to infinity. In this thesis, three different asymptotic settings are considered:

**Large-Dimensional Limit (LDL)** The LDL, \( n, p \to \infty, n/p \to c \), was already presented in the last section. It is the standard setting for the analysis of covariance matrices and in Random Matrix Theory in general (Bai & Silverstein, 2010). The LDL is also called the high-dimensional limit or the limit under Kolmogorov asymptotics. In the LDL, the sample covariance matrix is no longer a consistent estimator.

**Finite Observations Large-Dimensional Limit (FOLDL)** The FOLDL is an asymptotic setting introduced in Chapter 3 to make stronger statements on regularized estimators. In the FOLDL, the number of observations \( n \) is kept fixed and only the dimensionality tends to infinity. As for the LDL, the sample covariance is no longer consistent in the FOLDL.

To obtain theoretical results under LDL or FOLDL, I consider sequences of statistical models indexed by the dimensionality.

**General asymptotics (GA)** Under general asymptotics, \( p \) is seen as a function of \( n \), in such a way that \( p_n \) does not grow faster than linearly in \( n \) (Ledoit & Wolf, 2004). Note that GA includes standard asymptotics and the LDL as special cases. I focus on the LDL because (i) when \( p \) grows slower than linearly in \( n \), the sample covariance is known to be consistent in the limit and (ii) the clear relationship of \( p \) and \( n \) in the LDL allows for a better characterization of the dependency of regularized estimators on the structure of the population covariance matrix.
1.4.4 Evaluation of covariance estimators

The weaknesses of the sample covariance matrix described above are a strong motivation for the research on alternative estimators. To evaluate alternative estimators, it is necessary to define a metric which measures the quality of covariance matrix estimators. It is possible to distinguish between two general classes of quality metrics.

Measures of estimation error The first class of metrics are in the following called measures of estimation error. It contains all metrics which quantify the relationship of a covariance estimate or the distribution of a covariance estimate to the population covariance, the distribution of the data or the empirical distribution of the data. The negative data likelihood under Gaussianity, for example, relates an estimate to the empirical distribution of the data and its minimizer is the sample covariance matrix.

In this thesis, the most frequently used measure of estimation error is the expected squared error (ESE), for matrices also known as the expected Frobenius norm:

$$\Delta \hat{C} := E \left[ \sum_{ij} (\hat{C}_{ij} - C_{ij})^2 \right] = E\| \hat{C} - C \|^2$$

The percentage improvement in average loss (PRIAL) with respect to the sample estimator renders settings with different sample sizes and dimensionalities comparable:

$$\text{PRIAL}(\hat{C}) := 100 \cdot \frac{E\| \hat{S} - C \|^2 - E\| \hat{C} - C \|^2}{E\| \hat{S} - C \|^2}.$$ 

The PRIAL is a measure relative to the ESE of the sample estimator. A PRIAL of 100 means that the alternative estimator has no error while a PRIAL of 0 means that it yields no improvement. Negative values indicate performance worse than the sample covariance.

A variation is the PRIAL relative to the optimum in the sample eigenbasis. In the following, this measure is called the sample eigenbasis percentage improvement in average loss (SEPRIAL):

$$\text{SEPRIAL}(\hat{C}) := 100 \cdot \frac{E\| \hat{S} - S^* \|^2 - E\| \hat{C} - S^* \|^2}{E\| \hat{S} - S^* \|^2}.$$ 

Here $S^*$ is the covariance matrix with minimum mean squared error among those matrices with the same eigenvectors as the sample covariance matrix.

Other measures of estimation error considered in the literature are spectral norm (Cai et al., 2010), $L_1$-norm (Cai & Zhou, 2012a) and a scale-invariant loss function commonly referred to as Stein’s loss (James & Stein, 1961; Stein, 1986). A common alternative to the expected loss is minimax analysis (Dey & Srinivasan, 1985).

Finally, certain approaches are based on Riemannian Geometry which measure the geodesic distance (Smith, 2005; Barachant et al., 2012).

Measures of performance The second class of quality metrics is application-driven: in an application, a measure of estimation error such as ESE is not the quantity of interest, serving only as a proxy for an application-dependent performance measure:

- In a classification task, for example, the quantities of interest are classification accuracies, Receiver-Operator-Characteristics (ROC) or areas under ROC curves (AUC) (Fawcett, 2006). In this thesis, either plain accuracy or accuracy gain relative to usage of the sample covariance is reported:

  $$\text{accuracy gain}(\hat{C}) = \text{accuracy}(\hat{C}) - \text{accuracy}(\hat{S}).$$ (1.5)

---

3 With normalization n.
4 Some researchers consider the ESE of a triangular matrix to account for the symmetry of the covariance matrix.
• Another example considered in this thesis are portfolio performances, where the quantity of interest is a measure of financial risk.

It is difficult to say a priori whether a measure of estimation error is a good proxy for a specific performance measure. Yet, it is a question of high importance: the fact that most covariance estimators are (asymptotically) optimal for a specific loss function does not provide any guarantees for practical applications. As a consequence, it can be superior to optimize the regularization parameter of a regularized covariance estimator by cross-validation (Friedman, 1989). The downside of cross-validation is its high computational cost.

1.5 Approaches to covariance estimation

There exists a vast literature on covariance estimators. To clarify matters, I divide the estimators in three groups:

Structured estimators One way to overcome the large number of degrees of freedom is to restrict the space of positive definite matrices, analogous to restricting the function class in regression tasks. Various assumptions on the structure of the covariance matrix are possible, including sparsity, band-structure or low intrinsic dimensionality.

With respect to structured estimators, this thesis focuses on Factor Analysis (Basilevsky, 1994), a latent variable model for data with low intrinsic dimensionality, which is especially popular in finance. Structured estimators are discussed in Chapter 2.

Regularized estimators As in regression, an alternative to restricting the space of positive definite matrices is regularization. This reduces the effective degrees of freedom (Hastie et al., 2008). The most widely used type of regularization for covariance estimation is shrinkage (Stein, 1956; Friedman, 1989; Ledoit & Wolf, 2003b),

$$\tilde{C}^{shr}(\lambda) = (1 - \lambda)\hat{S} + \lambda\hat{T}, \quad (1.6)$$

where the shrinkage target $\hat{T}$ is a low-variance high-bias estimator of the covariance. Regularization introduces a bias and setting $\lambda$ can be seen as optimizing the trade-off between bias and variance (Stein, 1956). Regularized estimators are discussed in Chapter 3 and Chapter 4.

Estimators based on spectrum correction It was shown above that the distribution of the sample eigenvalues differs from the population eigenvalues. Spectrum correction approaches are used to infer a mapping from the sample eigenvalues to corrected eigenvalue estimates which yield a superior covariance matrix.

Spectrum correction dates back to Stein (Stein, 1986). Recently, the inversion of the Marčenko-Pastur Law led to exciting progress (El Karoui, 2008b). Spectrum correction is discussed in Chapter 5.

Other methods The field of covariance matrix estimation is quite large and highly specific estimators have been developed for specific applications. Some do not fit well into the groups defined above. These include work on the heteroskedasticity and autocorrelation consistent (HAC) estimation of covariances (Newey & West, 1987). In finance, covariance estimation is also needed for high-frequency data with asynchronous observations, which require distinct techniques (Hautsch et al., 2012). The field of robust statistics has also contributed to the research on covariance estimation by proposing estimators which are less sensitive to outliers (Huber, 1981).
1.6 Application domains contained in this thesis

Section 1.3 contained a literature overview of applications of covariance matrices. In the following chapters of this thesis, empirical results on four different data domains are presented: (i) daily stock market data, (ii) electroencephalography (EEG) data from Brain-Computer Interfaces (iii) recordings of spoken letters and (iv) images of digits.

1.6.1 Stock market data

Nowadays, the application of information technology to finance and stock markets is a widespread phenomenon. Fully electronic trading has become standard and, in some markets, trading algorithms execute more trades than human traders (Hendershott et al., 2011; Hendershott & Riordan, 2013).

History Even before electronic trading systems were put to use, information technology had a large impact on finance. The advent of modern finance began with Markowitz and his seminal paper on portfolio optimization (Markowitz, 1952). His theory provides a mathematical approach to diversification by directly minimizing the portfolio variance. Moreover, by adding constraints to the optimization problem, one can, e.g., prohibit short-selling. Other applications comprise the creation of portfolios which constitute optimal hedges or track indices. However, a fundamental issue in portfolio allocation is the accurate and precise estimation of the covariance matrix of asset returns from historical data.

Financial time series Since the publication of Markowitz, covariance estimation and coping with its uncertainties have occupied both researchers and practitioners in finance. The major difficulty with precise covariance matrix estimation in finance arises from the interplay between non-stationarity of financial time series and huge asset universes.

On the one hand, changes in the data generating processes force the estimation to rely on short time windows of recent observations (Loretan & Phillips, 1994; Pagan & Schwert, 1990).

On the other hand, modern stock markets permit the trading of more and more assets \( p \). The number of observations is linear in \( p \), while the degrees of freedom in the covariance matrix grows quadratically with \( p \). For example, in order to estimate the covariance matrix from the daily return series of a universe of one thousand assets, 500500 free parameters have to be estimated. Following a general rule of thumb, that 10 observations per parameter are required for a reliable estimate, the observation window would need to cover approximately twenty years of data. Such a temporal horizon, however, is at odds with reported nonstationarity of financial time series.

In practice, the situation is made even more complicated by the super-Gaussianity of financial time series (Loretan & Phillips, 1994; Longin, 2005; Campbell et al., 2008), which increases the difficulty of covariance estimation even further, especially in the case of small sample sizes.

Reuters tick data The data set considered in this thesis was aggregated from Reuters tick data. It consists of daily returns of about 1300 US stocks (3\(^{rd}\) Jan. 2001–2\(^{nd}\) Nov. 2009), about 600 European stocks (3\(^{rd}\) Jan. 2001–20\(^{th}\) April 2009) and a set of 200 stocks from the Hong Kong stock exchange (3\(^{rd}\) Jan. 2001–26\(^{th}\) Sept. 2008). Removing stocks which do not have data for the whole time horizon covered by the data set reduces the Hong Kong data set to 100 assets.

1.6.2 EEG data

Neuroscience and medical applications have advanced greatly through the measurement of neural activity. Since the second half of the 19\(^{th}\) century it has been known that electric currents form the basis of information processing. Berger (1929) first measured the resulting electric potential on the skin of the head and coined the term electroencephalography (EEG, for details see Kandel et al. (2000)).
EEG characteristics Listing some basic properties and comparing them to those of other neuroimaging techniques yields a characterization of EEG:
+ **Non-invasiveness**: EEG does not need surgery, which reduces health risks and ethical concerns.
+ **High temporal resolution**: compared to hemodynamic methods like functional magnetic resonance imaging (fMRI) and near infrared spectroscopy (NIRS) which are based on blood-oxygen-level dependent (BOLD) contrast, EEG can be recorded with very high temporal resolution (i.e. 1000Hz). In addition, the hemodynamic response suffers from a time delay of several seconds while EEG is instantaneous.
+ **Low cost**: the cost of an EEG system lies in a range which facilitates wide-spread clinical and commercial usage.
  - **Low signal-to-noise ratio (SNR)**: the membrane potential lies in the range of millivolts, yet on the scalp one measures a superposition of signals which is orders of magnitude weaker (in the range of microvolts). In addition, muscle and recording artifacts reduce the SNR (Fatourechi et al., 2007; Winkler et al., 2011).
  - **Low spatial resolution**: localizing the source of measured activity is a major issue in EEG and called the inverse problem (Baillet et al., 2001). **Spatial smearing**, caused by the low conductivity of the skull, makes the inverse problem difficult to solve (Nunez et al., 1997). In general, the challenges of neural data are similar to financial data: the data is non-stationary and non-Gaussian (McEwen & Anderson, 1975; von Bünau et al., 2009).

Brain-Computer Interfaces (BCI) A Brain-Computer Interface is a device which enables a human to gain control over a computer system relying solely on neural activity. The main motivation for this kind of research is to establish a communication pathway for ALS patients (e.g. Birbaumer et al. (1999); Müller et al. (2008); McCane et al. (2015)), but there exist further interesting applications such as gaming (Krepki et al., 2007; Nijholt et al., 2009; Marshall et al., 2013; Höhne et al., 2014).

There are applications of BCI technology without a feedback loop and there is an ongoing debate whether such a system should be termed BCI (Nijboer et al., 2013). Examples are the assessment of the vigilance of a car driver in mental state monitoring (MSM) BCIs (Hettinger et al., 2003; Blankertz et al., 2010b) or the assessment of (sub)conscious sensory perception in Neuro-Usability (NU) BCIs (Scholler et al., 2012; Porbadnigk et al., 2010, 2013). The following paragraphs present the two main neural signals exploited in BCI systems.

BCI based on motor imagery (MI) In an MI-based BCI, the subject communicates by imagining the movement of, e.g., the left or the right hand. The computer can then analyse the EEG time series to distinguish the activations (Pfurtscheller & Neuper, 2001). If successful, one bit of information has been transmitted.

In this thesis, the VITAL-BCI data set recorded at the TU Berlin and the University of Tübingen, first described by Blankertz et al. (2010a), is analyzed. After calibration, subjects used two imagined movements to control a 1D cursor application. For each of the 80 subjects, EEG activity during 150 feedback trials was recorded. For the analysis, \( p = 55 \) channels were selected and the time series was downsampled to \( n_{\text{trial}} = 390 \) measurements per trial.

BCI based on event-related potentials (ERP) ERP-based BCIs rely on potentials evoked by an exactly timed external sensory stimulus, where visual and auditory stimuli are most common.

In an ERP-BCI a sequence of \( k \) different stimuli is presented repetitively (\( m \) times) in a random order, where each stimulus is tied to an action in the BCI. The user attends to only one stimulus (target), while neglecting all others (non-targets). When the system correctly identifies the attended stimulus,
information has been transmitted. For each of the $k$ stimuli, the system evaluates the brain response at $m$ events and calculates a score reflecting the probability of whether or not the user was attending. Then, it takes a one-out-of-$k$-class decision based on the $k$ scores.

A detailed overview of state-of-the-art approaches for feature extraction and classification for ERP data in the BCI context is given in (Blankertz et al., 2011). In this thesis, a data set from an experiment comprising 21 subjects, using auditory stimuli (Schreuder et al., 2011), is analyzed. Sequentially, the subjects were presented with six different sounds coming from a range of speakers.

In a NU-ERP-BCI, the brain response to a stimulus is measured and the computer tries to infer the characteristics of the stimulus or its representation in the brain from the evoked potential. In this thesis, a data set which was recorded for a study in which 11 subjects had the task to distinguish between noisy and noise-free phonemes (Porbadnigk et al., 2010, 2013) is analyzed. For the data set, 427 standardized features were calculated from event-related potentials recorded in 61 electrodes. Ignoring noise intensity and discarding stimuli which the subjects did not identify, the trials are divided into two classes: correctly identified noise-free and correctly identified noisy phonemes.

1.6.3 Audio data: spoken letters

Speech recognition is a classical Machine Learning problem. In this thesis, covariance matrix estimators are evaluated on the ISOLET data set of spoken letters (Fanty & Cole, 1991) obtained from the UCI ML Repository (Bache & Lichman, 2013). For the data set, 617 empirically derived features such as Discrete Fourier Transform (DFT) coefficients and zero crossing rates were extracted from 7797 raw audio recordings of 26 letters. For the classifiers considered by Fanty & Cole (1991), the empirical features allow for better classification accuracies than the raw data.

It is important to point out that audio data is only used as a test bed for different covariance estimators. Approaches such as deep neural networks achieve higher performances on this kind of task (Hinton et al., 2012).

1.6.4 Image data: handwritten digits

Another classic application domain for Machine Learning algorithms is handwritten digit recognition. A standard data set is the United States Postal Service (USPS) handwritten digit database (Hull, 1994). It consists of 1100 observations with 256 pixels for each of the 10 digits. As for ISOLET, the USPS database is only used for evaluation of covariance estimators. Better classification rates for handwritten digits can be obtained with support vector machines (Schölkopf & Smola, 2002) and deep neural networks (Ciresan et al., 2012).

In addition, a similar data set is used, also obtained from the UCI ML Repository (Bache & Lichman, 2013), which consists of traces of handwritten digits (Alimoglu & Alpaydin, 1997). The special property of the data set is that each trace has, in addition to the class label, a subject label. This allows for consideration of transfer learning (see Section 4.1). The data set consists of 10992 traces, approximately equally distributed over 44 subjects and the 10 digits $0, 1, \ldots, 9$. The traces are converted into images of $30 \times 30$ pixels.

1.7 Contributions

Large parts of this thesis have been published in peer-reviewed journals or peer-reviewed conference proceedings. Here, I give a summary of the contributions in each of the articles.

1.7. CONTRIBUTIONS

The systematic error in Factor Analysis covariances is described and an algorithm for the correction is proposed and tested in simulations. Improved performance of portfolio allocations on three financial markets is demonstrated. The paper forms the basis of Chapter 2.


The importance of the covariance structure for the theory of analytic shrinkage is demonstrated. The theory is extended to arbitrary structures, properties of shrinkage are discussed and a shrinkage algorithm which performs better on four real world data sets is proposed. The paper is the basis for Section 3.2.


A new shrinkage algorithm for autocorrelated data is proposed which outperforms the existing approach in theory, simulations and on real world data. Shrinkage under autocorrelation is discussed in Section 3.3.


In a general framework, independent of a specific estimator, shrinkage is extended to multiple targets and conditions for consistency are derived. It is shown that the specific cases of mean and covariance fulfill the conditions and perform well in simulations and on real world data. The Multi-Target Shrinkage framework is presented in Chapter 4.


CVC, an alternative to Nonlinear Shrinkage based on cross-validation is proposed which is shown to have clear advantages. The CVC estimator is discussed in Section 5.
Chapter 2

Structured Estimation

The large number of degrees of freedom in high-dimensional covariance matrices makes their estimation difficult (see Section 1.4.2). Imposing restrictions on the covariance matrix is a straightforward way to reduce the degrees of freedom. This is also called *structured estimation* (Burg et al., 1982). The challenge lies in the selection of appropriate restrictions for a specific data domain. Popular structural assumptions on covariance matrices include sparsity, band-structure, and low intrinsic dimensionality.

**Sparsity** A covariance matrix in which most of the entries are zero is called *sparse*. Sparsity occurs when the majority of the $p$ variables are uncorrelated. There exists a large body of theoretical work on sparse covariance matrix estimation (Cai & Zhou, 2012b; El Karoui, 2008a; d’Aspremont et al., 2008; Lam & Fan, 2009), yet there are surprisingly few applications to real world data. Cai & Liu (2011) consider a gene microarray experiment, Bickel & Levina (2008a) analyze climate data.

A special case is sparse *inverse* covariance matrix estimation. The sparsity assumption in the inverse means that most variables are conditionally independent. Typical applications are genomics (Rothman et al., 2008) and proteomics (Friedman et al., 2008).

**Band-structure** A covariance matrix which has non-zero entries only close to the diagonal is said to have a *band-structure*. This may occur for spatial or temporal data, where it is assumed that data points with high spatial or temporal distance are uncorrelated (Cai et al., 2010; Cai & Yuan, 2012). *Tapering* replaces the hard threshold at a certain distance with a soft decay. Bickel & Levina (2008b) apply a band-structured covariance estimator to call-counts in a call center data set.

**Low intrinsic dimensionality** The perhaps most frequently applied structural assumption is *low intrinsic dimensionality*: for high-dimensional data, the dependencies between the variables are often given by a small number of latent components, also called factors, present in all variables. The best-known example for this is Spearman’s work on general intelligence: the performance of school children in a wide variety of subjects can be explained to a large part by a single latent factor which Spearman calls $g$ (Spearman, 1904).

In the following, the running example for low intrinsic dimensionality are stock market returns: the so-called *market factor* describes the overall behavior of the market. On a good day, nearly all stocks go up, on a bad day, nearly all stocks go down. This single factor already explains a large part of the variance of stock returns. Additional weaker factors exist, the exact number of relevant factors in stock market return data is an open question.

**Bias in factor models** This chapter focuses on covariance matrix estimation under the assumption of low intrinsic dimensionality. I show that covariance matrices derived from Factor Analysis exhibit
a systematic error, which is similar to the systematic error in the spectrum of the sample covariance matrix. I propose Directional Variance Adjustment (DVA), an algorithm which reduces the bias in Factor Analysis. A thorough empirical study for the US, European, and Hong Kong stock market shows that my proposed method leads to improved portfolio allocation and better reflects the market’s risk structure. The sections on the bias in Factor Analysis and the DVA approach are based on an article published in PLOS ONE (Bartz et al., 2013).

Chapter outline Section 2.1 introduces factor models and present Principal Component Analysis and Factor Analysis. Next, in Section 2.2, the bias in Factor Analysis is discussed, DVA is presented and the approach is illustrated in simulations. Section 2.3 applies DVA to portfolio optimization.

2.1 Factor models as structured covariance estimators

As discussed above, low intrinsic dimensionality means that a huge part of the variance is contained in a small number of components. This leads to the following definition of a factor model:

\[ x_t = W f_t + \epsilon_t \]  \hspace{1cm} (2.1)

\[ \epsilon_{it} \perp f_{jt}, \quad \forall i, j, \]  \hspace{1cm} (2.2)

\[ \epsilon_{it} \perp \epsilon_{jt}, \quad \forall i \neq j. \]  \hspace{1cm} (2.3)

The data time series \( x_t \) is generated by a linear combination of the \( M \) zero mean factors in the \((M \times 1)\)-vector \( f_t \), specified by a \((p \times M)\)-matrix of weights \( W \), and additive zero mean noise, given by the \((p \times 1)\)-vector \( \epsilon_t \). The factors and the noise are independent, as well as the noise of different dimensions (Fan et al., 2008; Goldfarb & Iyengar, 2003). In the statistics and signal processing literature, this is often referred to as an Independent Component Analysis (ICA) model, where \( W \) is the mixture matrix and \( f \) are the source signals (see, e. g., Hyvärinen & Oja (2000)).

Using the independence of factors and noise, eq. (2.2), the population covariance under the factor model is given by

\[ C_{FM} = \mathbb{E} [x_t x_t^T] = W \mathbb{E} [f_t f_t^T] W^T + \mathbb{E} [\epsilon_t \epsilon_t^T] \]  \hspace{1cm} (2.4)

where \( C_f \) is the covariance matrix of the factors and \( C_\epsilon \) is the covariance matrix of the noise. The factor covariance can be absorbed into the weights \( W \), hence without loss of generality one can assume \( C_f = I \). The noise covariance \( C_\epsilon \) is, by eq. (2.3), diagonal.

The advantage of this kind of model lies in the reduced number of degrees of freedom. As discussed in Section 1.2, an arbitrary covariance matrix \( C \) has \( \text{df}_C = \frac{1}{2} p(p + 1) \) parameters. To calculate the degrees of freedom in the factor model, one has to take the invariances in \( W \) into account (Akaike, 1987). This can be done by constraining the columns of \( W \) to be orthogonal, which yields

\[ \text{df}_{C_{FM}} = M \cdot p - \frac{M(M - 1)}{2} + \frac{p}{2} \]

\[ = (M + 1)p + M/2 - M^2/2. \]  \hspace{1cm} (2.5)

There is only a linear dependency on the dimensionality \( p \), hence for \( M \ll p \) there is a large reduction in the degrees of freedom. This reduction facilitates estimation by reducing the variance in the estimate. If the generative model does not follow the specified factor model, the structural assumption introduces a bias.

In general, optimizing the number of factors in a factor model is equivalent to optimizing the bias-variance trade-off: the smaller the number of factors, the larger the (potential) bias and the smaller the variance of the covariance estimate. In Section 2.1.3, some possibilities to determine the optimal number of factors are discussed.
2.1. FACTOR MODELS AS STRUCTURED COVARIANCE ESTIMATORS

2.1.1 Principal Component Analysis (PCA)

An obvious possibility to find these components is to search for the directions which maximize the sample variance:

$$\tilde{v}_i := \arg \max_{v} v^T \hat{S} v$$

$$\|v\|_2 = 1$$

The $\tilde{v}_i$ are the eigenvectors of the sample covariance matrix, with corresponding eigenvalues $\hat{\gamma}_i$. Based on the $M$ largest eigenvalues and corresponding eigenvectors, the principal components (Jolliffe, 2002), it is possible to construct a covariance estimator:

$$\hat{C}^{\text{PCA}} = \sum_{i=1}^{M} \hat{\gamma}_i \tilde{v}_i \tilde{v}_i^T.$$ 

This covariance matrix is rank deficient and hence of little use. By incorporating noise, full-rank matrices are obtained. There are two straightforward ways to do this:

- adding a multiple of the identity such that the trace of the sample covariance is conserved:

$$\hat{C}^{\text{PCA+isotropic noise}} := \hat{C}^{\text{PCA}} + \sigma^2 I,$$

$$\sigma^2 := \sum_i \hat{S}_{ii} - \sum_i \hat{C}_{ii}^{\text{PCA}}.$$ 

(2.6)

- adding a diagonal noise covariance matrix $\hat{C}^{\epsilon}$ such that the diagonal of the sample covariance is conserved:

$$\hat{C}^{\text{PCA+diag. noise}} := \hat{C}^{\text{PCA}} + \hat{C}^{\epsilon},$$

$$\hat{C}_{ii}^{\epsilon} := \hat{S}_{ii} - \hat{C}_{ii}^{\text{PCA}}.$$ 

(2.7)

The case with isotropic noise is also called a scalar factor model (Connor et al., 2010). Both of these approaches are used in practice. This is again a case of bias-variance trade-off: the model with isotropic noise has lower variance at the cost of (potentially) higher bias than the more flexible diagonal noise model. This makes the isotropic noise model favorable when (i) there are very few data points or (ii) the distribution of the data set under consideration is well described by isotropic noise. The case (ii) is given, for example, when data is standardized (Laloux et al., 2000).

2.1.2 Probabilistic PCA and Factor Analysis

In the above factor models based on principal components, the definition of the factors and the integration of the additional noise was presented as a heuristic. As an alternative, it is possible to specify a fully probabilistic generative model by assuming Gaussian distributed factors $f_t$ and noise $\epsilon_t$ (cf. Figure 2.1),

$$f_t \sim \mathcal{N}(0, I), \quad \epsilon_t \sim \mathcal{N}(0, C^{\epsilon}),$$

and search for the “best” explanation of the observed data for a given number of factors. Here, the “best” model refers to the model that maximizes the data (log-)likelihood:

$$L(W, C^{\epsilon}) = \ln \int_F p(X, F|W, C^{\epsilon}) dF.$$ 

It has been shown that for isotropic noise $C^{\epsilon} = \sigma^2 I$, weights in $W$ correspond to the eigenvectors and that the covariance estimate is equal to eq. (2.6). Hence the model with isotropic noise is named Probabilistic PCA (PPCA, Tipping & Bishop (1999)).
The case of an arbitrary diagonal noise matrix $C^e$ is called Factor Analysis (FA, see Basilevsky (1994)). For FA, the optimal factors differ from the principal components. In other words: with respect to the likelihood, the model eq. (2.7) is suboptimal.

Maximum likelihood estimation in such latent variable models is difficult due to the marginalization over factors. The standard approach to this problem is Expectation-Maximization (EM, see Dempster et al. (1977)), for applications to Factor Analysis see Rubin & Thayer (1982); Roweis & Ghahramani (1999)). In this algorithm, the likelihood is maximized iteratively by alternating between the Expectation and the Maximization step:

- in the Expectation step, the weights $W$ and noise variances $C^e$ are assumed to be fixed and the expected factors $F$ (latent variables) can be derived directly.

- in the Maximization step, the expected factors $F$ are assumed to be fixed and the likelihood is maximized with respect to exposures $W$ and noise variances $C^e$.

These two steps are iterated until convergence.$^1$

**Further factor models** There exists a large number of different models and extensions. Instead of finding components which maximize variance or likelihood, it is also possible to search for independent components (ICA, see Comon (1994); Hyvärinen & Oja (2000)). A popular extension of PCA is Kernel Principal Component Analysis (KPCA, see Schölkopf et al. (1998)).

$^1$Alternative approaches to solving the optimization problem are proposed in the literature, the best known is the quasi-newton method by Jöreskog (1967). As that algorithm uses an eigendecomposition, which is costly to obtain in high dimensions ($O(N^3)$), the EM approach ($O(MN^2)$) is used in the following. Other methods claiming superior performance suffer from the same drawback (see, e.g., Zhao et al. (2008)). Moreover, for the discussion of the systematic error, the optimization procedure chosen to obtain the maximum likelihood solution is of no importance.
2.1.3 Model selection for the number of factors

The number of factors is an important hyperparameter of all types of factor models. A large set of model selection approaches has been proposed in the literature, a discussion in detail is beyond the scope of this thesis.

- **Information criteria**: the standard methods are based on the Akaike Information Criterion (AIC, see Akaike (1987)), Bayesian Information Criteria (BIC, see Schwarz et al. (1978)) or related Information criteria (Lopes & West, 2004).

- **Bayesian treatment**: a Bayesian treatment of factor models allows to automatically determine an effective number of components: setting the number of factors to $p-1$ and putting a hierarchical prior over the weights $W$, certain directions in space can be switched off (Bishop, 1999). This is related to the framework of automatic relevance determination (see MacKay (1995)).

- **Random Matrix Theory**: it is possible to estimate the number of factors using Random Matrix Theory. Laloux et al. (2000) and Rosenow et al. (2002) propose to choose the number of factors in a PCA factor model such that only those eigenvalues are retained which are larger than the largest eigenvalues of a random matrix, given by the Marchenko-Pastur law. These factors are likely to reflect some real structure.

- **Cross-validation**: As for (nearly) all model selection problems, cross-validation is a reliable, but computationally costly, alternative.

In the following, the model selection issue is neglected.

2.2 Directional Variance Adjustment

As there is no analytic solution for the parameters of the Factor Analysis model, I cannot provide a stringent theoretic analysis of its properties. Instead, by means of thorough simulations, evidence is provided that the spectrum of the covariance matrix derived from a Factor Analysis model is biased and an algorithm for bias reduction is proposed.

2.2.1 Systematic error in Factor Analysis

A result comparable to the Marchenko-Pastur law, which would relate the population factor model to the maximum likelihood solution, does not exist for Factor Analysis. Therefore, I run a simulation to study systematic errors in Factor Analysis.

**Measuring bias of directional variances**  Figure 1.2 illustrated the systematic error of the eigen-spectrum of the sample covariance matrix by comparing population and sample eigenvalues. These are defined by the population and sample variances in the population and sample eigendirections $v_i$ and $\hat{v}_i$, respectively:

$$\gamma_i := v_i^T C v_i \quad \text{vs.} \quad \hat{\gamma}_i := \hat{v}_i^T \hat{S} \hat{v}_i.$$

Because of the differences between the $v_i$ and the $\hat{v}_i$, this comparison does not capture the entire estimation error. In the following, a set of directions is defined which is tailored to the estimate of the Factor Analysis model. Then, systematic errors, defined by directional variances based on population variances and estimated variances, are studied.

More precisely, systematic errors in the factor subspace and its complementary orthogonal space are investigated separately. To this end, an orthonormal basis $\hat{P}_{fs}$ ($p \times M$) of the $M$-dimensional subspace, containing the estimated factors $\hat{W}$ (the **Factor Subspace**), and another orthonormal basis $\hat{P}_{oc}$ ($p \times (p-M)$)
of the \((p - M)\)-dimensional orthogonal complement, are defined. Correspondingly, the covariance matrix is projected to the two subspaces, yielding a factor space related part and its orthogonal counterpart as

\[
\begin{align*}
\hat{C}_{fs}^{FA} &= \hat{P}_{fs}^{0} \hat{P}_{fs}^{0 \top} \hat{C}_{fs}^{FA} \hat{P}_{fs}^{0} \hat{P}_{fs}^{0 \top}, \\
\hat{C}_{oc}^{FA} &= \hat{P}_{oc}^{0} \hat{P}_{oc}^{0 \top} \hat{C}_{FA} \hat{P}_{oc}^{0} \hat{P}_{oc}^{0 \top}.
\end{align*}
\]

From the eigenvectors of \(\hat{C}_{fs}^{FA}\) and \(\hat{C}_{oc}^{FA}\) for non-zero eigenvalues, two new bases, \(\hat{P}_{fs}\) and \(\hat{P}_{oc}\), are obtained. The eigenvectors are assumed to be sorted in decreasing order with respect to the eigenvalues. Combining these subspace bases to \(\hat{P} = [\hat{P}_{fs}, \hat{P}_{oc}]\) yields an orthonormal basis of the entire space.

Along these directions \(\hat{p}_i\), the directional variances \(\sigma^2_i\) of the population factor model and the estimated Factor Analysis model are measured and the systematic error is calculated as

\[
\eta_i^{FA} := E \left[ \frac{\hat{\sigma}^2_{i, FA}}{\sigma^2_i} \right], \quad \hat{\sigma}^2_i := \hat{p}_i^{\top} \hat{C}_{FM} \hat{p}_i, \quad \sigma^2_{i, FA} := \hat{p}_i^{\top} \hat{C}_{FA} \hat{p}_i.
\]

Here, values \(\eta_i^{FA} > 1\) and \(\eta_i^{FA} < 1\) correspond to an over- and underestimation of the directional variances, respectively. Moreover, the basis \(\hat{P}\) explicitly takes the factor structure into account. Hence, this particular basis enables the study of the specific systematic estimation errors in the factor subspace and noise subspace separately. Note that the directions \(\hat{p}_i\) are solely derived from the estimated parameters of the factor model and do not rely on information about the population covariance or the generative model.

**Illustration of the bias** For the simulation, artificial Gaussian data \((p = 30)\) is generated according to an underlying three factor model as in eq. (2.1). The noise covariance matrix \(C^e\) is defined with equally spaced values from the interval \([0.5, 1.5]\) on the diagonal. The three columns of the mixing matrix \(W\) are generated as randomly oriented vectors with a length of 10, 3 and 1, respectively. In order to study the small sample size properties of Factor Analysis for this setting, the ratio \(n/p\) is set to 0.7, 1 and 5, corresponding to 21, 30, and 150 thirty-dimensional observations. As \(W\) and \(C^e\) are known for the simulation, the population covariance matrix \(C_{FM}\) is given by eq. (2.4).

Figure 2.2 depicts the estimated systematic error \(\hat{\eta}^{FA}_i\) of Factor Analysis as defined in eq. (2.10) by means of the simulated data. Clearly, Factor Analysis tends to overestimate the variance in the 3-dimensional Factor Subspace, while the variance in the orthogonal complement is on average underestimated. This is not surprising, as the Factor Analysis model attributes strong covariances in the sample to the factors. Consequently, factors with low Signal-to-Noise-ratio (SNR) are hard to identify and directions of spurious covariance are likely to be misrepresented as factors, leading to an overestimation of the variance along these directions: in the simulations, the strongest (first) factor, which has a high Signal-to-Noise-Ratio, can be estimated with very high accuracy even for small sample sizes and the variance estimate does not have a significant systematic error. The weaker factors with a lower SNR in contrast are not identified well, FA tends to yield overestimated variances along the estimated factor directions. This effect is highly pronounced for small sample sizes and persists for relatively large sample sizes.

On the other hand, the noise subspace spectrum shows a similar — albeit weaker — behavior as the spectrum of the sample covariance matrix, i.e., variances corresponding to large eigenvalues are overestimated, while variances corresponding to small eigenvalues are underestimated (compare Figure 1.2 and Figure 2.2). As for the sample covariance matrix, this effect is especially pronounced for small sample sizes.

### 2.2.2 Directional Variance Adjustment: correcting the systematic error

The correction of the systematic error is more involved for the FA covariance estimate than for the sample covariance matrix: the minimization of the Factor Analysis cost function does not have a closed-form solution. Instead, it requires an iterative method. This does not facilitate an analytical approach
2.2. DIRECTIONAL VARIANCE ADJUSTMENT

Figure 2.2. Systematic error of Factor Analysis. Average ratio between Factor Analysis and population variances \( \hat{\eta} \) in the factor subspace and the orthogonal complement. Ratios of sample size to dimensionality \( n/p = 0.7, 1 \) and 5. \( p = 30 \). Average over 150 simulations.

To obtain the distribution of the eigenvalues. Consequently, I deploy a method based on a parametric bootstrap (Casella & Berger, 2002).\(^2\)

Algorithm description Suppose we have estimated the parameters \( \mathcal{F} = (\mathbf{W}, \mathbf{C}^r) \) of a Factor Analysis model and want to correct the corresponding covariance matrix \( \hat{\mathbf{C}}_{\text{FA}(\mathcal{F})} \) for the systematic error. Then it is possible to estimate the systematic error in the following way: using \( \mathcal{F} \) for a generative model, we generate \( R_{mc} \) synthetic data sets of the same size as the original sample. For each data set we estimate a corresponding Factor Analysis parameter set \( \mathcal{F}_1, \ldots, \mathcal{F}_{R_{mc}} \). Note that for these parameter sets the ground truth (i.e., \( \mathcal{F} \)) is known and with it the population covariance matrix. This enables us to quantify the amount by which the directional variances along the eigendirections of \( \hat{\mathbf{C}}_{\text{FA}(\mathcal{F}_r)} \) (factor subspace, eq. (2.8)) and \( \hat{\mathbf{C}}_{\text{oc}(\mathcal{F}_r)} \) (orthogonal complement, eq. (2.9)) are over- and underestimated, respectively. Then, the estimated systematic errors can be directly turned into multiplicative correction factors for the adjustment of the directional variances of \( \mathcal{F} \). Applying these corrections to the eigendirections of the factor space and its orthogonal complement yields what is in the following referred to as the directional variance adjusted covariance matrix estimate \( \hat{\mathbf{C}}_{\text{DVA}(\mathcal{F})} \) (see Algorithm 1).

Note that the algorithm does not correct the parameters of the factor model itself. Instead, only the resulting covariance matrix is adjusted. In particular, the factor directions are kept unchanged.

Illustration of the algorithm Figure 2.3 illustrates the adjustment of the covariance matrix. The figure shows the covariances of the population and the estimated factor model in blue and red, respectively. The arrows indicate the directions of the factors and the orthogonal complement. Clearly, the factor direction is subject to estimation error and its strength is overestimated. The variance in the direction orthogonal to the factor is underestimated. The proposed DVA method corrects the systematic

\(^2\)An alternative to the parametric bootstrap, it is possible to deploy cross-validation to directly estimate the directional variances, in a manner similar to the CVC estimator proposed in Section 5.3.
Algorithm 1 DVA

**Input:** the estimated parameters of the Factor Analysis model $\mathcal{F}$; the sample size $n$; the number of Monte Carlo runs $R_{mc}$

**Output:** the directional variance adjusted covariance estimate $\hat{C}^{\text{DVA}(\mathcal{F})}$

1. generate $R_{mc}$ synthetic data sets of size $n$ based on $\mathcal{F}$.
2. from the $R_{mc}$ data sets, estimate $R_{mc}$ factor model parameter sets $\mathcal{F}_1, \ldots, \mathcal{F}_{R_{mc}}$
3. For each $\mathcal{F}_r$, estimate the basis $\hat{P}_r = [\hat{P}_{r,fs}, \hat{P}_{r,oc}]$
4. estimate the directional variance correction factors
   $$\hat{\eta}_{ri}^{\text{corr}} = \frac{1}{R_{mc}} \sum_{r=1}^{R_{mc}} \frac{\hat{P}_{r,fs}^\top \hat{P}_{r,fs} \hat{P}_{r,fs}^\top \hat{P}_{r,oc} \hat{P}_{r,oc}}{n \hat{P}_{r,fs}^\top \hat{P}_{r,oc}}$$
5. For $\mathcal{F}$, estimate the basis $\hat{P} = [\hat{P}_{fs}, \hat{P}_{oc}]$
6. calculate the directional variance adjusted covariance matrix
   $$\hat{C}^{\text{DVA}(\mathcal{F})} = \hat{C}^{\text{FA}(\mathcal{F})} + \sum_{i=1}^{N} (1 - 1/\hat{\eta}_{ri}^{\text{corr}}) \cdot (\hat{p}_i \hat{p}_i^\top) \hat{C}^{\text{FA}(\mathcal{F})} (\hat{p}_i \hat{p}_i^\top)$$

error in the estimate of the directional variances along these directions, without adjusting the directions themselves. This leads to the directional variance adjusted covariance matrix (depicted in green): in the aforementioned directions, the systematic error is reduced.

**Limitations of the parametric bootstrap** One has to keep in mind that the resampling—and with it the estimate of the systematic error of the FA covariance matrix—is based on the estimated parameters $\mathcal{F}$. Therefore, large errors in $\mathcal{F}$ adversely affect the DVA covariance estimate.

In order to reduce the impact of the error in $\mathcal{F}$, it could be advantageous to iterate the DVA procedure. From the DVA covariance matrix, which more closely reflects the population covariance matrix, we could estimate the parameters of a new factor model and restart the DVA procedure, obtaining more precise estimates of correction factors in each iteration. Though a compelling idea, there is neither a guarantee that iterating the DVA method yields a better solution, nor that it converges to a sensible one or even converges at all. In the following, only the non-iterated DVA procedure is considered.

### 2.2.3 Simulation results

In this section, the effectiveness of the proposed DVA method is illustrated in a simulation study. For this, standard and DVA FA are compared on toy data, generated according to the scheme presented in section 2.2.1.

The performances of the two estimation methods with respect to the systematic error $\eta_i^{\text{FA}}$ (eq. (2.10)) are contrasted in Figure 2.4. To the left, it is shown that the DVA method clearly reduces the systematic error of the Factor Analysis model, even for relatively large ratios $n/p$. In the direction of the third factor, which has the lowest SNR, the reduction is most salient. In the orthogonal complement of the factor subspace, the adjusted spectrum is a very good estimate of the population variances. Nevertheless, there remains a small systematic error, which is due to using the estimated parameter set in order to infer the directional variance correction factors. The right panel of Figure 2.4 illustrates that the DVA method does not suffer from a relevant increase in variance of the estimate.

By reducing the systematic error with almost no increase in variance, the DVA method reduces the average estimation error. To account for different magnitudes of population directional variances, Figure 2.5 displays the error of the estimator in terms of the mean absolute relative error

$$\xi_i^{\text{FA/DVA}} := \mathbb{E} \left[ \left| \frac{\hat{\sigma}_{i,\text{FA/DVA}}^2 - \hat{\sigma}_{i,\text{FA}}^2}{\hat{\sigma}_{i,\text{FA}}^2} \right| \right].$$  (2.11)
2.3 Evaluation in a portfolio simulation

Figure 2.3. Illustration of the DVA algorithm. Depicted are directional variances for the estimated (red) and population Factor model covariance matrix (blue). The blue squares indicate population variances along the estimated factor direction and the direction of the orthogonal complement. The DVA method (green) aims at stretching and compressing the estimated covariance matrix such that the variance estimates in these directions correspond to the population variances.

Note that this error is more than halved for the direction of the low SNR-factor and considerably decreased in the orthogonal complement. Here, DVA has the strongest effects on the directions corresponding to the largest and smallest non-zero eigenvalues of $C^{FA}_{oc}$. For the direction of the smallest eigenvalue, the error is again approximately halved.

While the ratio $n/p$ determines most of the properties of the sample covariance, this is not true for regularized estimators and factor models. For larger values of $n$, at a constant ratio $n/p$, the noise variances of Factor Analysis are estimated more precisely, while the estimation of the factors remains difficult. This is shown in Figure 2.6, where the dimensionality has been set to 500 and the generative model has seven factors of strength 10, 5, 4, 3, 2.5, 2, 1.5, and 1. One can see that while there is little room for improvement in the orthogonal complement, in the factor subspace the performance gain by DVA remains on the same level as in the simulations with smaller dimensionality.

2.3 Evaluation in a portfolio simulation

In order to evaluate the proposed method, DVA Factor Analysis is applied to a financial data set of daily return time series. In the experiments, covariance matrices of stock returns are estimated and then used for portfolio optimization. The realized risks of the portfolios are compared for different covariance estimators. For the sake of comparison, approaches which do not first estimate a covariance matrix and...
Figure 2.4. Comparison of the systematic error in standard Factor Analysis and DVA Factor Analysis. Left: systematic error. Right: normalized standard deviation of the error. Simulations for different ratios of sample size to dimensionality (n/p = 0.7, 1 and 5). p = 30. Correction factors estimated on R = 100 generated data sets. Mean over 150 simulations.

Figure 2.5. Comparison of the mean absolute relative error for standard Factor Analysis and DVA Factor Analysis for different ratios of sample size to dimensionality (n/p = 0.7, 1 and 5). p = 30. Correction factors estimated on K = 100 generated data sets. Averaged over 150 simulations.

then use that covariance matrix to optimize portfolios are included.

2.3.1 Approaches compared in the simulations

The DVA Factor Analysis method is compared to the following competing approaches:

- Sample covariance matrix: the standard baseline is given by the sample covariance matrix.
2.3. EVALUATION IN A PORTFOLIO SIMULATION

Figure 2.6. Comparison of the mean absolute relative error for standard Factor Analysis and the DVA Factor Analysis in high dimensions for different ratios of sample size to dimensionality (n/p = 0.7, 1 and 5). Note that the y-axis has different scaling for the factor subspace and the orthogonal complement. p = 500. Correction factors estimated on K = 100 generated data sets. Mean over 150 simulations.

- **Equal weights portfolios:** the second baseline method does not require a covariance matrix: the investment is spread evenly over all assets (DeMiguel et al., 2009). Alternatively, this can be seen as an optimization based on a covariance matrix restricted to the class of all multiples of the identity.

- **Resampling Efficiency:** Monte Carlo sampling, as in DVA, is known in the portfolio optimization literature from Resampling Efficiency (Michaud, 1998). There, the authors follow a fundamentally different approach. While DVA uses resampling to reduce the bias of the factor model, in Resampling Efficiency the sample mean and covariance are used to generate additional data sets, on which optimal portfolio weights are calculated and then averaged. This is supposed to lead to more stable and diversified portfolios, but there is an ongoing debate about the merits of this procedure (Scherer, 2004). Resampling Efficiency does not yield a covariance estimate. Though not based on Monte Carlo sampling, techniques for the correction of variance inflation in principal components analysis are more related to the DVA algorithm (Kjems et al., 2001; Abrahamsen & Hansen, 2011).

- **Fama-French three-factor model:** Factor Analysis is a purely statistical model. A set of models which incorporate additional information is given by fundamental factor models (Connor, 1995; Connor et al., 2010). In a fundamental factor model, assets are analyzed and certain key metrics are used to set up the factor model. This makes them especially well suited when only a short history of data is available, e.g. for weekly or monthly data, as fewer parameters need to be estimated from the history than in a statistical factor model.\(^3\)

\(^3\)In contrast, macroeconomic factor models predetermine the factors as macroeconomic time series which are supposed to affect the asset returns. As in the Fama-French model, the exposures are obtained by linear regression. Examples for macroeconomic time series used in factor models are unemployment rate, GNP, FX or interest rates. However, for daily or higher frequency stock market returns, macroeconomic factor models are of limited use and therefore neglected in the following (for an overview, see Connor et al. (2010)).
The best-known model of this kind is the Fama-French three-factor model, for which the factor time series $f$ are based on portfolios governed by market beta (exposure to general market movements), book-to-market ratio and market capitalization (Fama & French, 1992). The exposures to these factors are obtained from the coefficients of a linear regression model.

- **Ledoit-Wolf Shrinkage**: a regularized estimator is given by Ledoit-Wolf Shrinkage, discussed in detail in the next chapter. In finance, a (fundamental) one-factor model has proven to be the best choice of shrinkage target (Ledoit & Wolf, 2003a,b). To differentiate from the standard target, this approach is called LW-FMT-Shrinkage in the following.

- **Standard Factor Analysis**: the application of Factor Analysis to financial data was first introduced in order to test the Arbitrage Pricing Theory and is well-established (Roll & Ross, 1980). Probabilistic PCA is not flexible enough to capture the largely different variances of different stocks.

For DVA and standard Factor Analysis seven factors are used. Though on the higher dimensional US and EU data sets it is possible to extract more meaningful factors while fewer factors are favorable on the smaller HK data set, opting for the same intermediate model complexity on all data sets keeps the setting simpler. In fact, the reduced estimation bias makes DVA Factor Analysis less sensitive to the inclusion of too many factors.

### 2.3.2 Design of portfolio simulations

There are various applications of covariance matrices in portfolio optimization. Covariance matrices are needed for index tracking, hedging and the minimization of portfolio variance. In the following, the focus lies on minimum variance portfolios,

$$\omega^* = \arg \min_{\omega} \omega^\top \bar{C} \omega, \quad (2.12)$$

where $\omega$ is the vector of portfolio weights and $\bar{C}$ is an estimate of the covariance matrix.

**Constrained portfolio optimization** Depending on the particular application, additional constraints are incorporated into the optimization. Commonly applied constraints include:

- $\sum_i \omega_i = 1$: the sum of all portfolio weights is restricted to one.
- $\omega^\top \bar{x} = \bar{r}$: the estimated portfolio return is restricted to $\bar{r}$, where $\bar{x}$ is the vector of predicted asset returns.
- $\omega_i \geq 0$: only positive portfolio weights (a negative weight on an asset corresponds to short-selling).

Note that the placing of constraints tremendously reduces the set of feasible portfolios and hence diminishes the influence of the covariance estimate (Jagannathan & Ma, 2003). Consequently, the observed differences between the performances of portfolios obtained from different covariance estimation methods become smaller. Thus, in order to examine the advantages of the various covariance estimation methods, no positivity or magnitude constraint was put on the weights. The only imposed constraint is that the portfolio weights sum to one\textsuperscript{4}.

\textsuperscript{4}This optimization is independent of the return estimates and is equivalent to optimizing portfolio returns under the assumption of equal expected returns for all assets. A discussion of the structure of expected returns (see, e.g. $\beta$-pricing models, Shanken (1992)) is beyond the scope of this thesis.
2.3. EVALUATION IN A PORTFOLIO SIMULATION

Out-of-sample evaluation In order to evaluate the performance of the different covariance estimators, the risk of the estimated portfolios is measured with two quantities:

- **Realized (out-of-sample) variance** of the portfolios:
  \[
  \sigma^2_{\text{real}} = \frac{1}{n} \sum_{t=1}^{n} \left[ \omega_{t-1}^\top (x_t - \hat{x}_{t-1}) \right]^2.
  \] (2.13)

- Of more financial interest is the **realized mean absolute deviation**:  
  \[
  \text{MAD}_{\text{real}} = \frac{1}{n} \sum_{t=1}^{n} \left| \omega_{t-1}^\top (x_t - \hat{x}_{t-1}) \right|.
  \] (2.14)

Two additional quantities are typically considered in portfolio optimization studies:

- The **Sharpe-Ratio** is a risk-adjusted performance measure of an investment. It is defined by
  \[
  S := \frac{\mathbb{E} \left[ \omega_{t-1}^\top x_t - r_b \right]}{\text{Var}(\omega_{t-1}^\top x_t - r_b)},
  \]
  where \( r_b \) is the baseline of a risk-free (low return) investment. For daily returns, the risk-free return is approximately zero.

- The **turnover** quantifies the stability of portfolios over time, which is important given that every change of the portfolio is associated with trading costs. It is defined by
  \[
  t := \sum_{t=1}^{n} \sum_{i=1}^{p} \left| \omega_{p,t} - \omega_{p,t-1} \right|.
  \]

Note that (2.13) and (2.14) are rolling out-of-sample estimates, as \( \hat{x}_{t-1} \) and \( \omega_{t-1} \) are estimates of the expected return and optimal portfolio weights based on the information available until time \( t - 1 \). More precisely, a strictly causal window of 150 trading days is chosen for covariance matrix estimate \( \hat{C}_{t-1} \) and the return estimate \( \hat{r}_{t-1} \). This size balances out two error sources:

- Covariance estimation gets less precise for fewer data points.
- Non-stationarity of the return distribution limits the usefulness of older data.

In this thesis, slightly more observations are used compared to studies with monthly data (Ledoit & Wolf, 2003b; Jagannathan & Ma, 2003), but less than e.g. Bouchaud & Potters (2011) use for daily data. For the considered data sets, larger estimation windows increase portfolio variance.

In order to reduce the variance of the performance evaluation and to thoroughly explore the estimated covariance structure, \( J = 1000 \) random subsets, each confined to 40 (HK) or 100 (US and EU) assets, are chosen and the optimal (confined) portfolio \( \omega_j^t \) is constructed from the given covariance matrix estimate \( \hat{C}_j^t \). The realized variance and realized absolute deviation are then determined based on the average performance across the different confined portfolios, i.e.,

- \[
  \sigma^2_{\text{real}} = \frac{1}{n} \sum_{t=1}^{n} \left\{ \frac{1}{J} \sum_{j=1}^{J} \left[ \omega_{j,t-1}^\top (x_t - \hat{x}_{t-1}) \right]^2 \right\}.
  \]
- \[
  \text{MAD}_{\text{real}} = \frac{1}{n} \sum_{t=1}^{n} \left\{ \frac{1}{J} \sum_{j=1}^{J} \left| \omega_{j,t-1}^\top (x_t - \hat{x}_{t-1}) \right| \right\}.
  \]
**Additional regularization**  Minimization of eq. (2.12) only yields the optimal portfolio for the population covariance matrix. Portfolio optimization based on covariance estimates tends to put too much weight on directions of small empirical covariance. As a consequence, a regularized optimization problem which enforces stronger diversification can yield portfolios with reduced risk.

Assuming independence, the optimal portfolio allocation $\omega_{\text{indep}}$ has weights inverse to the variance of the assets; it is highly diversified. To gradually interpolate between $\omega^*$ and $\omega_{\text{indep}}$, a new set of weights is defined,

$$
\omega_{\text{reg}}(\lambda) := \arg \min_{\omega} \omega^\top C \omega + \lambda \omega^\top M \omega,
$$

(2.15)

where the metric $M$ is set to a diagonal matrix which has the single asset variances on its diagonal. Then, each asset gets penalized by its variance and in the limit $\lambda \to \infty$ the portfolio $\omega_{\text{indep}}$ is obtained.

The aim of the following analysis is twofold: first of all and from a theoretical perspective, it is of interest to investigate whether it is possible to explain away the superior performance of the DVA method by a higher degree of diversification or whether the population covariance structure is captured more accurately. Second, with respect to practical considerations, it is desired to achieve the best performance possible.

**2.3.3 Portfolio simulation results**

The results for the different markets are summarized in Table 3.2. First note that the equal weights portfolio has very high risk, a result which is also reported by Kourtis et al. (2012). As expected, the sample covariance matrix is ill-suited for portfolio optimization: across all data sets, the portfolios derived from the various factor based models and shrinkage clearly outperform the sample covariance matrix based portfolios in terms of realized risk.

A direct comparison of the structured and regularized models reveals that the DVA method always significantly outperforms Fama-French, standard Factor Analysis and shrinkage with respect to realized variance and realized absolute deviation. On the data sets considered, Resampling Efficiency does not yield an advantage over the sample covariance matrix.

**Sharpe-Ratio and turnover**  The Sharpe-Ratios do not give a clear picture: Fama-French, statistical factor modeling and shrinkage each perform best in one market. This unclear result is not surprising, as portfolios were not optimized for high returns. The turnovers show an additional advantage of DVA Factor Analysis over standard Factor Analysis: DVA yields covariance estimates of higher stability and thereby reduces trading costs.

**Results under additional regularization**  Figure 2.7–2.9 depict the realized (out-of-sample) variance and MAD (see eq. (2.13) and eq. (2.14)) of the resulting portfolios as a function of the regularization parameter $\lambda$ (eq. (2.15)) for each of the three market samples. Equal weights portfolios have been omitted in the figures because they incur far higher risk.

Each of the models benefits from additional regularization, as can be seen from a reduction of the realized risk of the resulting portfolios (see Tables 3.2 and 2.2). Although the improvement is most pronounced for the sample covariance matrix, it merely reaches the performance of unregularized Factor Analysis\(^5\). Again, Resampling Efficiency is not superior to the sample covariance matrix.

LW-FMT-Shrinkage, which minimizes the expected squared error (ESE), also profits from additional regularization to $M$. This indicates that the ESE is not a good proxy for portfolio performance: the regularization which yields minimal ESE is weaker than the regularization which yields optimal portfolios.

Surprisingly, Fama-French (i) benefits less from regularization than LW-FMT-Shrinkage, although the unregularized performance is similar, and (ii) regularization does not reduce the performance difference

\(^5\)Note that the regularized optimization based on the sample covariance matrix is equivalent to unregularized optimization using a shrinkage covariance estimator, that employs $T = M$ as the shrinkage target (see eq. (1.6))
2.3. Evaluation in a Portfolio Simulation

Table 2.1. Portfolio risk. Mean absolute deviations $10^3$, mean squared deviations $10^6$, Sharpe-Ratio and turnover of the resulting portfolios for the different covariance estimators and the different markets. † := DVA mean significantly better/worse than this model at the 5% level, tested by a randomization test.

<table>
<thead>
<tr>
<th></th>
<th>MAD</th>
<th>MSE</th>
<th>Sharpe</th>
<th>turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>US</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>equal weights</td>
<td>10.63t</td>
<td>246.8t</td>
<td>0.61</td>
<td><strong>0.00t</strong></td>
</tr>
<tr>
<td>Sample Covariance</td>
<td>8.56t</td>
<td>156.1t</td>
<td>0.51t</td>
<td>6.49t</td>
</tr>
<tr>
<td>Resampling Efficiency</td>
<td>8.83t</td>
<td>165.6t</td>
<td>0.50t</td>
<td>6.81t</td>
</tr>
<tr>
<td>Fama-French</td>
<td>5.65t</td>
<td>73.5t</td>
<td><strong>0.77</strong></td>
<td>2.06t</td>
</tr>
<tr>
<td>LW-FMT-Shrinkage</td>
<td>5.56t</td>
<td>69.6t</td>
<td>0.74t</td>
<td>2.69t</td>
</tr>
<tr>
<td>Factor Analysis</td>
<td>5.47t</td>
<td>67.8t</td>
<td>0.73t</td>
<td>2.53t</td>
</tr>
<tr>
<td>DVA Factor Analysis</td>
<td><strong>5.40</strong></td>
<td><strong>66.7</strong></td>
<td>0.72</td>
<td>2.33</td>
</tr>
<tr>
<td><strong>Europe</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>equal weights</td>
<td>8.10t</td>
<td>154.8t</td>
<td>0.65t</td>
<td><strong>0.00t</strong></td>
</tr>
<tr>
<td>Sample Covariance</td>
<td>5.93t</td>
<td>78.9t</td>
<td>0.91t</td>
<td>5.09t</td>
</tr>
<tr>
<td>Resampling Efficiency</td>
<td>6.11t</td>
<td>83.4t</td>
<td>0.89t</td>
<td>5.32t</td>
</tr>
<tr>
<td>Fama-French</td>
<td>3.97t</td>
<td>38.6t</td>
<td>1.16t</td>
<td>1.71t</td>
</tr>
<tr>
<td>LW-FMT-Shrinkage</td>
<td>4.00t</td>
<td>39.1t</td>
<td>1.26t</td>
<td>2.24t</td>
</tr>
<tr>
<td>Factor Analysis</td>
<td>3.88t</td>
<td>36.5t</td>
<td><strong>1.30t</strong></td>
<td>2.06t</td>
</tr>
<tr>
<td>DVA Factor Analysis</td>
<td><strong>3.84</strong></td>
<td><strong>36.0</strong></td>
<td>1.29</td>
<td>1.91</td>
</tr>
<tr>
<td><strong>Hong Kong</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>equal weights</td>
<td>10.21t</td>
<td>209.5t</td>
<td>1.17t</td>
<td><strong>0.00t</strong></td>
</tr>
<tr>
<td>Sample Covariance</td>
<td>6.57t</td>
<td>81.2t</td>
<td>1.22t</td>
<td>2.06t</td>
</tr>
<tr>
<td>Resampling Efficiency</td>
<td>6.64t</td>
<td>82.7t</td>
<td>1.21t</td>
<td>2.13t</td>
</tr>
<tr>
<td>Fama-French</td>
<td>6.20t</td>
<td>73.5t</td>
<td><strong>1.37</strong></td>
<td>1.46t</td>
</tr>
<tr>
<td>LW-FMT-Shrinkage</td>
<td>6.17t</td>
<td>72.9t</td>
<td>1.37t</td>
<td>1.62t</td>
</tr>
<tr>
<td>Factor Analysis</td>
<td>6.17t</td>
<td>73.0t</td>
<td>1.37t</td>
<td>1.64t</td>
</tr>
<tr>
<td>DVA Factor Analysis</td>
<td><strong>6.12</strong></td>
<td><strong>71.7</strong></td>
<td>1.37</td>
<td>1.52</td>
</tr>
</tbody>
</table>

Figure 2.7. Regularization dependency of the realized portfolio risk in the US market, measured in mean absolute deviation (left) and variance (right).
to the statistical factor models FA and DVA FA. This indicates that the unregularized Fama-French model achieves good performance because it yields strongly diversified portfolios, but that the covariance structure is better captured by the statistical factor models.

Regularization also considerably reduces the portfolio risk for Factor Analysis and DVA FA. Optimally regularizing each method, the DVA FA model significantly outperforms the sample covariance and Fama-French for all markets. The performance gain over shrinkage is, however, not significant for the HK market.

For optimal regularization the difference in performance between standard Factor Analysis and DVA Factor Analysis is reduced. This was expected, as regularization is similar to constraining the feasible set and hence also diminishes the influence of the covariance estimate on the minimum variance portfolio (Jagannathan & Ma, 2003). Thus regularization partly compensates for the influence of the systematic error of the Factor Analysis covariance matrix estimate.

Nevertheless, in the US and EU market, the performance gain in MAD of DVA over standard Factor
2.3. EVALUATION IN A PORTFOLIO SIMULATION

Table 2.2. Portfolio risk under regularization. Mean absolute deviations $10^3$, mean squared deviations $10^6$, Sharpe-Ratio and turnover of the resulting portfolios for the different regularized covariance estimators for optimal regularization strength and the different markets. $\dagger := \text{DVA mean significantly better/worse than this model at the 5\% level, tested by a randomization test.}$

<table>
<thead>
<tr>
<th>Market</th>
<th>MAD</th>
<th>MSE</th>
<th>Sharpe</th>
<th>turnover</th>
</tr>
</thead>
<tbody>
<tr>
<td>US</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>equal weights</td>
<td>9.78$\dagger$</td>
<td>212.9$\dagger$</td>
<td>0.61</td>
<td>0.00$\dagger$</td>
</tr>
<tr>
<td>Sample Covariance</td>
<td>5.45$\dagger$</td>
<td>67.3$\dagger$</td>
<td>0.76</td>
<td>1.90</td>
</tr>
<tr>
<td>Resampling Efficiency</td>
<td>5.48$\dagger$</td>
<td>67.7</td>
<td>0.76</td>
<td>1.81</td>
</tr>
<tr>
<td>Fama-French</td>
<td>5.55$\dagger$</td>
<td>70.0$\dagger$</td>
<td>0.77</td>
<td>1.61</td>
</tr>
<tr>
<td>LW-FMT-Shrinkage</td>
<td>5.39$\dagger$</td>
<td>65.8</td>
<td>0.75</td>
<td>1.73</td>
</tr>
<tr>
<td>Factor Analysis</td>
<td>5.38$\dagger$</td>
<td>66.0</td>
<td>0.74</td>
<td>1.75</td>
</tr>
<tr>
<td>DVA Factor Analysis</td>
<td><strong>5.35</strong></td>
<td><strong>65.6</strong></td>
<td><strong>0.72</strong></td>
<td><strong>1.65</strong></td>
</tr>
<tr>
<td>Europe</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>equal weights</td>
<td>7.50$\dagger$</td>
<td>133.6$\dagger$</td>
<td>0.66$\dagger$</td>
<td><strong>0.00$\dagger$</strong></td>
</tr>
<tr>
<td>Sample Covariance</td>
<td>3.91$\dagger$</td>
<td>37.0</td>
<td>1.26$\dagger$</td>
<td>1.69</td>
</tr>
<tr>
<td>Resampling Efficiency</td>
<td>3.93$\dagger$</td>
<td>37.2</td>
<td>1.26</td>
<td>1.61</td>
</tr>
<tr>
<td>Fama-French</td>
<td>3.93$\dagger$</td>
<td>37.7</td>
<td>1.17</td>
<td>1.44</td>
</tr>
<tr>
<td>LW-FMT-Shrinkage</td>
<td>3.86$\dagger$</td>
<td>36.3</td>
<td>1.28</td>
<td>1.53</td>
</tr>
<tr>
<td>Factor Analysis</td>
<td>3.82$\dagger$</td>
<td>35.6</td>
<td><strong>1.30</strong></td>
<td>1.56</td>
</tr>
<tr>
<td>DVA Factor Analysis</td>
<td><strong>3.81</strong></td>
<td><strong>35.5</strong></td>
<td>1.29</td>
<td>1.50</td>
</tr>
<tr>
<td>Hong Kong</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>equal weights</td>
<td>9.50$\dagger$</td>
<td>181.7$\dagger$</td>
<td>1.19$\dagger$</td>
<td><strong>0.00$\dagger$</strong></td>
</tr>
<tr>
<td>Sample Covariance</td>
<td>6.14$\dagger$</td>
<td>72.9</td>
<td>1.48</td>
<td>1.10</td>
</tr>
<tr>
<td>Resampling Efficiency</td>
<td>6.16$\dagger$</td>
<td>73.4$\dagger$</td>
<td>1.48</td>
<td>1.09</td>
</tr>
<tr>
<td>Fama-French</td>
<td>6.11</td>
<td>71.7</td>
<td>1.50</td>
<td>1.02</td>
</tr>
<tr>
<td>LW-FMT-Shrinkage</td>
<td>6.10</td>
<td>71.8</td>
<td><strong>1.51</strong></td>
<td>1.06</td>
</tr>
<tr>
<td>Factor Analysis</td>
<td>6.09</td>
<td>71.7</td>
<td>1.50</td>
<td>1.07</td>
</tr>
<tr>
<td>DVA Factor Analysis</td>
<td><strong>6.09</strong></td>
<td><strong>71.3</strong></td>
<td>1.49</td>
<td>1.03</td>
</tr>
</tbody>
</table>

Analysis remains significant at the 5\% level. In Hong Kong the difference is (for optimal regularization) not significant.

Compared to the American or European market, the Hong Kong market shows a slightly different behavior. At the Hong Kong market, all methods likewise benefit from additional diversification. One possible explanation is that the HK market data contains many outliers and missing data as opposed to the US and EU market data. Thus covariance estimates as well as least square estimates of factor exposures are less precise. For the HK market data, in contrast to US and EU, the Fama-French model clearly profits from additional regularization. Nevertheless, its overall performance remains inferior to DVA Factor Analysis.

2.3.4 Conclusions from portfolio simulations

The fundamental issue in portfolio allocation is the accurate and precise estimation of the covariance matrix of asset returns from historical data. Among many challenges, the data is typically high-dimensional and contaminated with outliers, and its non-stationarity interferes with the use of long estimation windows. Thus, reliable statistical parameter estimation is impeded.

In the previous sections, I (i) demonstrated that the data-driven statistical Factor Analysis model has a systematic estimation error, (ii) proposed the algorithmic Directional Variance Adjustment (DVA) framework, which alleviates this bias, and finally (iii) provided extensive simulations concerning minimum
variance portfolios of EU, US and Hong Kong markets, underpinning the usefulness of the DVA approach in terms of significant gains in realized variance and realized mean absolute deviation. Compared to standard Factor Analysis, covariance estimates are more stable and turnover is reduced.

**Conclusions from the regularization experiment**  For each covariance estimator, I also studied the effect of regularizing the minimum variance portfolios towards a higher degree of diversification. As expected, diversification improved portfolio performance across the different estimators. The empirical study showed that while regularization slightly decreases the overall advantage gained by DVA, the remaining difference in the minimum remained significant for the US and EU data sets, where the DVA Factor Analysis method proved to be superior to standard Factor Analysis.

The regularization experiments yielded surprising results for the Fama-French model: its advantage over the sample covariance matrix estimator can be attributed to an imposed strong diversification prior rather than to an improved estimation of the underlying covariance structure. The combination of regularization with statistical factor models such as standard FA and in particular DVA FA led to clearly better model performance.

It is possible to interpret the regularization in eq. (2.15) as a modification of the covariance estimate:

\[
\hat{C}_{\text{mod.}} := \hat{C} + \lambda M.
\]

While each estimate \(\hat{C}\) is optimal w.r.t. a measure of estimation error like ESE or likelihood, this optimality no longer holds for the modified estimates (\(\lambda > 0\)). Nevertheless, for suitable \(\lambda\) the modified estimates yield superior portfolio performance. This shows that the measures of estimation error are not perfect proxies for the portfolio performance.

**Additional aspects**  Normality is a strong assumption, especially in the finance context. In order to render Factor Analysis more suitable for financial data, an extension to t-distributions (t-FA) is possible (McLachlan et al., 2007). As t-FA has the same kind of bias as standard FA, DVA can be applied without modifications.

Hybrid factor models combine statistical, fundamental and/or macroeconomic factors (Connor, 1995; Miller, 2006; Connor et al., 2010). Hybridization was not considered in the simulations, but as long as a hybrid model contains statistical factors, the DVA approach can be applied to improve covariance estimation.

### 2.4 Chapter summary

Restricting the space of positive semi-definite matrices by imposing structure reduces the degrees of freedom and allows for more precise covariance matrix estimation. In this chapter, I have focused on the assumption of low intrinsic dimensionality. The corresponding structured estimators are factor models, for which the correlations between observations are completely modeled by a small number of factors.

Assuming that (i) there is a factor structure in the data and (ii) the number of factors has been specified correctly, covariance estimates based on factor models have less systematic error than the sample covariance. Yet, a systematic error remains. For factor models based on Principal Components Analysis, this bias can be described and corrected using tools from Random Matrix Theory.

For Factor Analysis, the systematic error has not been discussed before. I have shown in simulations that the systematic error can be quite substantial and proposed the DVA algorithm which yields covariance estimates with reduced systematic error.

Extensive portfolio simulations have shown that the DVA covariance estimates better reflect the risk structure than competing approaches and consistently yield lower portfolio risks.
Chapter 3

Regularized Estimation

The previous chapter discussed estimators which restrict the model class of covariance matrices. In this chapter, regularized covariance estimators are introduced. Regularization can be seen as a reduction of the number of effective degrees of freedom (Hastie et al., 2008). Yet, it is perhaps more intuitive to see regularization as a way to reduce the variance of an estimator at the cost of introducing a bias.

Effective regularization requires two choices. First, the type of regularization has to be chosen in a way which reflects the assumptions on the population covariance matrix. This is similar to linear regression, where an $L_1$ or an $L_2$ penalty is chosen depending on whether the true solution is expected to be sparse or dense in the weights, respectively (Hastie et al., 2008). Second, the trade-off between bias and variance has to be optimized (Stein, 1956). Regularization to the extreme yields a large bias, while insufficient regularization does not reduce the variance sufficiently.

Shrinkage

This chapter focuses on regularization approaches which linearly combine the sample covariance matrix with a biased low-variance estimator (Haff, 1980). This is related to shrinkage, a technique for the improved estimation of the sample mean, first proposed by Stein (1956). In covariance shrinkage, one considers the convex combination of the sample covariance matrix $\mathbf{S}$ and a shrinkage target $\mathbf{T}$,

$$\mathbf{C}^{shr}(\lambda) = (1 - \lambda)\mathbf{S} + \lambda\mathbf{T}, \quad (3.1)$$

which is a low-variance high-bias estimator of the covariance. The advantage of this approach is that there is only one free parameter, the shrinkage intensity $\lambda$, which can be estimated via cross-validation (CV-Shrinkage) or an analytic formula (Ledoit & Wolf, 2004). The analytic formula, Ledoit-Wolf Shrinkage (LW-Shrinkage), is consistent in the large-dimensional limit (LDL).

Shrinkage and covariance structure

I show that the proof of consistency requires bounds on the growth rates of the eigenvalues and their dispersion and demonstrate that these bounds are often violated for real world data. Without imposing restrictions on the covariance structure and therefore better accommodating real world data, I provide a detailed analysis in the LDL. I prove consistency and derive conditions under which consistency even holds in the FOLDL. The analysis shows that in most relevant cases the shrinkage intensity converges to zero. This motivates an extension of LW-Shrinkage – automatic orthogonal complement Shrinkage – which adapts to the covariance structure. The corresponding sections in this chapter are based on a article published in the NIPS 2013 proceedings (Bartz & Müller, 2013).

Shrinkage and autocorrelation

LW-Shrinkage assumes i.i.d. data, a fact which is often ignored in the literature. In the presence of autocorrelated data, shrinkage intensities have a negative bias and the minimum of the expected squared error (ESE) is no longer achieved. The Sancetta-Estimator is an extension of LW-Shrinkage beyond i.i.d. data which is consistent in the LDL. I show that the Sancetta-Estimator
suffers from a large bias when dealing with finite sample sizes and propose an improved bias-corrected autocorrelation consistent shrinkage estimator which is (i) unbiased, (ii) less sensitive to hyperparameter choice and (iii) yields superior performance in simulations and on real world data. These findings and the proposed estimator were published in the NIPS 2014 proceedings (Bartz & Müller, 2014).

Chapter outline The chapter begins with an introduction to shrinkage in Section 3.1, where I also present the mathematical details of LW-Shrinkage and give an overview over the literature. Section 3.2 deals with the influence of the covariance structure on LW-Shrinkage. In Sections 3.2.1-3.2.4 the implied assumptions in LW-Shrinkage and their violation in real world data are discussed. In Section 3.2.5, aoc-Shrinkage is proposed, whose superior performance on real world data is demonstrated in Section 3.2.6.

Section 3.3 analyzes shrinkage under autocorrelation. In Section 3.3.1, I illustrate the bias introduced by a violation of the i.i.d. assumption, present the Sancetta-Estimator and propose an improved bias-corrected autocorrelation consistent shrinkage estimator. Sections 3.3.2-3.3.4 show the superiority of the proposed estimator in theory, simulations and on real world data.

3.1 Introduction to analytic shrinkage

The concept of shrinkage dates back to Charles Stein who proved the surprising result that the multivariate sample mean is not an admissible estimator of the population mean. An estimator is only admissible if there exists no estimator which is better for all possible values of the population parameter. In particular, Stein showed that “shrinking” to a constant target vector dominates the sample mean (Stein, 1956; James & Stein, 1961). Later, he showed that the sample variance is also inadmissible (Stein, 1964). This section considers the extension of the shrinkage concept to covariance matrices, which has become popular recently (Ledoit & Wolf, 2004).

The shrinkage covariance estimator is defined in terms of the convex combination\footnote{Arbitrary linear combinations are also possible, see Ledoit & Wolf (2004).} eq. (3.1) of sample covariance $\hat{S}$ and shrinkage target $\hat{T}$, the default being:

$$\hat{T}^{\text{id}} := \hat{\nu} I, \quad \hat{\nu} := p^{-1} \text{Tr}(\hat{S}),$$

where $\text{Tr}(\hat{S})$ is the trace of the sample covariance matrix. For an appropriate shrinkage intensity, this estimator has a set of favorable properties:

(i) reduced bias in the spectrum: the over-estimated large eigenvalues become smaller, the underestimated small eigenvalues become larger.

(ii) improved stability of matrix inversion: regularizing towards $\hat{T}^{\text{id}}$ with condition number equal to one, the condition number of the shrinkage estimator is smaller than the condition number of the sample covariance matrix and hence the inversion is more stable.

(iii) smaller ESE than the sample covariance: Ledoit & Wolf (2004) showed that for the optimal shrinkage intensity, the shrinkage estimator is guaranteed to have a smaller ESE than the sample covariance matrix.

Estimating the optimal shrinkage intensity The standard procedure for setting the shrinkage intensity is cross-validation (CV), which yields estimates of the out-of-sample performance based on hold-out sets (Friedman, 1989). The drawback of this CV-Shrinkage is its high computational cost: for (i) problems with many hyperparameters, (ii) very high-dimensional data sets, or (iii) online settings which need fast responses, CV-Shrinkage can become infeasible and a faster model selection method is required.
3.1. INTRODUCTION TO ANALYTIC SHRINKAGE

Haff (1980) proposed the first analytic alternative, based on an empirical Bayesian approach:

\[ \hat{C}^{\text{Haff}} = \frac{pn - 2n - 2}{pn^2} m_{\text{EB}} I + \frac{n}{n + 1} \hat{S}, \]

where \( m_{\text{EB}} = |\det(\hat{S})|^{1/p} \). More recently, Ledoit & Wolf (2004) derived an analytic formula for the regularization parameter \( \lambda \) in eq. (3.1) which minimizes the expected mean squared error of the convex combination.

LW-Shrinkage has become widely used mainly due to three properties. It has (i) appealing consistency properties in the limit of infinitely many observations and dimensions, (ii) negligible computational cost when used in high dimensions for algorithms and applications which rely on expensive matrix inversions or eigendecompositions and (iii) yields very accurate covariance estimates, as Ledoit & Wolf (2004) showed in extensive simulations.

**Derivation of LW-Shrinkage** The analytic shrinkage formula results from directly minimizing the expected squared error,

\[ \Delta \hat{C}^{\text{shr}} (\lambda) := \mathbb{E} \left[ \| C - (1 - \lambda) \hat{S} - \lambda \hat{T} \|^2 \right], \quad (3.3) \]

given by the squared distance between the population covariance matrix \( C \) and the shrinkage covariance matrix eq. (1.6):

\[
\lambda^* := \arg \min_{\lambda} \mathbb{E} \left[ \| C - (1 - \lambda) \hat{S} - \lambda \hat{T} \|^2 \right] \\
= \arg \min_{\lambda} \sum_{i,j} \left\{ 2\lambda \left[ \text{Cov}(\hat{S}_{ij}, \hat{T}_{ij}) - \text{Var}(\hat{S}_{ij}) \right] + \lambda^2 \mathbb{E} \left[ (\hat{S}_{ij} - \hat{T}_{ij})^2 \right] + \mathbb{E} \left[ \text{Var}(\hat{S}_{ij}) \right] \right\} \quad (3.4) \\
= \frac{\sum_{i,j} \left[ \text{Var}(\hat{S}_{ij}) - \text{Cov}(\hat{S}_{ij}, \hat{T}_{ij}) \right]}{\sum_{i,j} \mathbb{E} \left[ (\hat{S}_{ij} - \hat{T}_{ij})^2 \right]} \quad (3.5)
\]

The analytic estimate of the optimal shrinkage intensity \( \hat{\lambda} \) is obtained by replacing expectations with sample estimates:

\[
\hat{\text{Var}}(\hat{S}_{ij}) := \frac{1}{(n-1)n} \sum_s \left( x_{is}x_{js} - \frac{1}{n} \sum_t x_{it}x_{jt} \right)^2 \quad (3.6)
\]

\[
\hat{\text{Cov}}(\hat{S}_{ii}, \hat{T}_{ii}) := \frac{1}{(n-1)n p} \sum_k \left\{ \sum_s x_{is}^2 x_{ks}^2 - \frac{1}{n} \sum_t x_{it}^2 \sum_u x_{iu}^2 \right\}
\]

\[
\hat{\mathbb{E}} \left[ (\hat{S}_{ij} - \hat{T}_{ij})^2 \right] := (\hat{S}_{ij} - \hat{T}_{ij})^2,
\]

where \( \hat{\text{Cov}}(\hat{S}_{ii}, \hat{T}_{ii}) \) is shown for \( \hat{T} = \hat{T}^{\text{id}} \). These estimators assume \( \mathbb{E}[x_{it}] = 0 \) which greatly simplifies the notation; the extension to arbitrary means is straightforward.

**Theoretical analysis in Ledoit & Wolf (2004)** Under standard asymptotics, \( p \) fixed while \( n \) goes to infinity, the sample covariance matrix is a consistent estimator: this implies that the shrinkage intensity is zero in the limit. Shrinkage is designed for high-dimensional problems in which the number of observations is of the same order of magnitude as the number of variables. Ledoit & Wolf (2004) showed that the estimator \( \hat{\lambda} \) is consistent under general asymptotics (see Section 1.4.2).
This theoretical result is based on the analysis of a sequence of statistical models indexed by \( n \). \( X^n \) denotes a \( p_n \times n \) matrix of \( n \) i.i.d. observations of \( p_n \) variables with mean zero and covariance matrix \( C^n \). \( Y^n = R^n X^n \) denotes the same observations in their eigenbasis with diagonal covariance \( \Gamma^n = R^n C^n R^n \). The \( \gamma_i^n \) denote the eigenvalues of \( C^n \) or \( \Gamma^n \). Lower case letters \( x_i^n \) and \( y_i^n \) denote the entries of \( X^n \) and \( Y^n \), respectively.

Note that the estimation of shrinkage intensities is invariant to rotations. Switching to the eigenbasis, in which the sample covariance is denoted by \( S_0 \), often simplifies the analysis.

Assumptions in Ledoit & Wolf (2004) To prove consistency, Ledoit & Wolf (2004) make the following assumptions:

\[
\begin{align*}
\text{(LW1)} & \quad \exists K_1 \forall n : p_n/n \leq K_1, \\
\text{(LW2)} & \quad \exists K_2 : \frac{1}{p_n} \sum_{i=1}^{p_n} \mathbb{E}[(y_i^n)^2] \leq K_2, \\
\text{(LW3)} & \quad \lim_{n \to \infty} \frac{p_n^2}{n^2} \sum_{i,j,k,l \in Q_n} \left( \text{Cov}(y_i^n y_j^n, y_k^n y_l^n) \right)^2 = 0.
\end{align*}
\]

Here, \( Q_n \) is the set of all quadruples consisting of distinct integers between 1 and \( n \).

Assumption (LW1) just defines the framework of general asymptotics described above. The second assumption (LW2) is not discussed by Ledoit & Wolf (2004). The third assumption (LW3) states that, in the limit, the average of covariances between the products of uncorrelated variables is zero. As Ledoit & Wolf (2004) explain, this assumption is relatively weak and holds, amongst others, for elliptically distributed data. Assumption (LW3) is needed because a corresponding term—which convergence to zero does not follow from the other assumptions—appears in the consistency proof.

Shrinkage target options Many alternatives to the standard target exist. Schäfer & Strimmer (2005) proposed a whole set of shrinkage targets, including:

- \( \bar{T}_{\text{diag}} = \hat{S} \circ I \) (elementwise product)
- \( \bar{T}_{\text{const. corr.}} = \hat{S} \circ I + \bar{G}_{\text{cc}} \circ (1 \cdot 1^T - I) \),
  where \( \bar{G}_{ij} = \sqrt{\hat{S}_{ii} \hat{S}_{jj}} \cdot \bar{p} \) and \( \bar{p} \) is the average correlation between dimensions.
- \( \bar{T}_{\text{perf. corr.}} = \hat{S} \circ I + \bar{G}_{\text{pc}} \circ (1 \cdot 1^T - I) \),
  where \( \bar{G}_{ij} = \sqrt{\hat{S}_{ii} \hat{S}_{jj}} \).

These targets can be seen as restricted estimators. In fact, the standard target is a zero-factor PPCA and the diagonal target a zero-factor FA covariance. For financial data, Ledoit & Wolf (2003b) propose to use a 1-factor model as the shrinkage target.

The structure of the population covariance matrix determines which of these structured estimators is optimal. In practice, the choice is based on expert knowledge or cross-validation. An alternative is proposed in Chapter 4: Multi-Target Shrinkage optimizes over multiple target candidates at the same time and yields even lower estimation error than the best individual shrinkage target.

Extensions and variants There exists a wide variety of shrinkage estimators, of which I can only give a small overview. As many applications require the inversion of the covariance matrix, Kourtis et al. (2012) proposed direct shrinkage of the inverse. To deal with dependent data, Sancetta (2008) generalized the estimator to incorporate time-lagged observations of the covariance (see Section 3.3.1).

\(^2\)The sequence index \( n \) and the observation index \( t \) are often dropped to improve readability of formulas.
3.2. SHRINKAGE FOR ARBITRARY COVARIANCE STRUCTURES

The LW-Shrinkage framework makes few assumptions on the distribution of the data. Imposing stronger assumptions, it is possible to improve upon the estimator described above: Chen et al. (2010) proposed alternative algorithms, Rao-Blackwell-Ledoit-Wolf Shrinkage and Oracle Approximating Shrinkage, which yield (slightly) more accurate estimates of $\lambda^*$ under Gaussianity.

### 3.2 Shrinkage for arbitrary covariance structures

Above, the assumptions (LW1)-(LW3) were stated which are required for consistency under general asymptotics, and it was discussed that (LW1) and (LW3) are not problematic. In this section, it is shown that this does not hold for (LW2), the bound on the eighth moment.

Sancetta (2008) observed that (LW2) implies a growth rate on the Frobenius norm of the covariance matrix. Yet, the role of (LW2) has not been profoundly analyzed. In particular, the applicability of LW-Shrinkage to real world data has simply been taken for granted (Ledoit & Wolf, 2003b; Schäfer & Strimmer, 2005; Blankertz et al., 2011; Lemm et al., 2011). In the following, the interplay between structural assumptions on the covariance matrix and LW-Shrinkage is discussed in detail. I contribute in four aspects:

First, I derive three intuitive and easily testable necessary conditions for assumption (LW2) and show that many real world data sets of practical relevance violate these conditions.

Second, I design assumptions which better fit the statistical properties observed in many real world data sets, not posing restrictions on the covariance structure. I prove that LW-Shrinkage remains consistent under these new assumptions. In addition, I show that it is possible to consistently estimate the shrinkage intensity in the FOLDL.

Third, I provide a detailed analysis of the asymptotic behavior of LW-Shrinkage, which yields a counter-intuitive result: for typical covariance structures, no shrinkage—and therefore no regularization—takes place in the LDL. In practice, this leads to small shrinkage intensities and degrading performance.

Finally, I propose an extension of the shrinkage framework which for typical covariance structures does not suffer from zero shrinkage in the limit: automatic orthogonal complement Shrinkage (aoc-Shrinkage) takes the structure into account and outperforms LW-Shrinkage on real world data at the cost of a moderate increase in computation time.

#### 3.2.1 Implicit assumptions on the covariance structure

This section provides important insight into the implications of the assumption (LW2). The assumption (LW2) bounds the average size of the eighth moments in the eigenbasis $y_n^{i1}$. This is a stronger assumption than bounding the eighth moments of the observations $x_n^{i1}$ because it implies assumptions on the covariance structure. The following theorems quantify this relationship and provide conditions which are easily tested on empirical data.

**Theorem 3.1** (Largest eigenvalue growth rate). Let (LW2) hold. Then, there exists a limit on the growth rate of the largest eigenvalue

$$\gamma_1^n := \max_i \text{Var}(y_n^i) = O\left(p_n^{1/4}\right).$$

This is the smallest rate which bounds all sequences.

**Proof.** See Appendix B.1. □

Theorem 3.1 shows that (LW2) implies a bound on the largest eigenvalue. The consistency proof of LW-Shrinkage only holds for covariance matrices which fulfill this bound.
Theorem 3.2 (Dispersion growth rate). Let (LW2) hold. Then, there exists a limit on the growth rate of the normalized eigenvalue dispersion

\[ d_n := \frac{1}{p_n} \sum_i \left( \gamma_i - \frac{1}{p_n} \sum_j \gamma_j \right)^2 = O(1). \]

This is the smallest rate which bounds all sequences.

Proof. See Appendix B.2.

Theorem 3.2 shows that (LW2) imposes an even stronger restriction on the eigenvalue structure: the normalized dispersion must be bounded, which rules out that a set of eigenvalues of increasing set size grows at a smaller rate than defined by Theorem 3.1.

Theorem 3.3 (Sparsity in the LDL). Let \( \varepsilon > 0, \alpha > 0, \) and \( r := \frac{|W_n|}{p_n} \) with \( W_n := \{(i,j) \in \mathbb{C}_{ij}^n \mid |C_{ij}| > \varepsilon \} \) the set of all index pairs for which the entry in the covariance matrix is larger than \( \varepsilon \). If (LW2) holds and

\[ \lim_{n \to \infty} p_n = \infty, \]

then

\[ \lim_{n \to \infty} r = 0. \]

Proof. See Appendix B.3.

Theorem 3.3 tells us that the restrictions on the asymptotic case are quite severe. Setting an arbitrarily small threshold \( \varepsilon \), in the limit nearly all \( p_n \) entries in the covariance matrix have to be smaller than this threshold to fulfill (LW2). From the theorem it can be seen that the number of entries which are allowed to exceed this threshold is proportional to \( p_n \). As \( \varepsilon \) can be chosen arbitrarily small, this means that (LW2) is only fulfilled for matrices which are sparse in the limit, e.g. for block-diagonal matrices with fixed block size or band-structured matrices with finite and fixed bandwidth. The special structure of these matrices permits estimation with superior convergence rates than the estimation of general covariance matrices (Cai et al., 2010; Cai & Zhou, 2012b; Cai & Yuan, 2012).

Illustration in simulation  Theorems 3.1-3.3 show that (LW2) restricts the covariance structure of the sequence of models as the dimensionality increases. To illustrate this, two sequences of models A and B of increasing dimensionality \( p \) are generated. In A and B, the variables \( x_i^p \) have different probabilities of being correlated with a signal \( s^p \):

\[ x_i^p = \begin{cases} (0.5 + b_i^p) \cdot \varepsilon_i^p + \alpha \varepsilon_i^p s^p, & \text{with probability } P_{s_{A/B}}(i), \\ (0.5 + b_i^p) \cdot \varepsilon_i^p, & \text{else}. \end{cases} \tag{3.7} \]

Here \( b_i^p \) and \( \varepsilon_i^p \) are uniform random from \([0,1]\), \( s^p \) and \( \varepsilon_i^p \) are standard normal, \( \alpha = 1 \), \( P_{s_{A/B}}(i) = 0.2 \) and \( P_{s_A}(i) = (i/10 + 1)^{-7/8} \) (power law decay). To avoid systematic errors, the ratio of observations to dimensions is kept fixed: \( n^p/p = 2 \).

Covariance matrices from model A (Figure 3.1, left) are dense in the upper left corner, corresponding to the first few dimensions. With growing dimensionality, the matrix becomes sparse and thereby fulfills the requirement of (LW2) shown in Theorem 3.3. In contrast, covariance matrices of model B display evenly spread correlations.

The two plots to the right in Figure 3.1 depict normalized sample dispersion and the largest eigenvalue. For model A, one observes the behavior postulated by Theorems 3.1 and 3.2: the dispersion is bounded, the largest eigenvalue grows with the fourth root. For model B, there is a linear dependency of both dispersion and largest eigenvalue: (LW2) is violated.
3.2. SHRINKAGE FOR ARBITRARY COVARIANCE STRUCTURES

Figure 3.1. Simulated covariance matrices and dependencies of the largest eigenvalue/dispersion on the dimensionality. Average over 100 repetitions.

Figure 3.2. Empirical dependency of the largest eigenvalue/dispersion on the dimensionality. Average over 100 random subsets.

Empirical validation For real world data, the dependency of the largest eigenvalue/dispersion on the dimensionality is measured by averaging over random subsets. Figure 3.2 shows the results for four data sets presented in Section 1.6: (i) New York Stock Exchange, (ii) USPS hand-written digits, (iii) ISOLET spoken letters and (iv) the NU-ERP-BCI data set.

For all four data sets, the largest eigenvalues and the normalized dispersions (see Figure 3.2) closely resemble model B; a linear dependence on the dimensionality which violates (LW2) is visible.

3.2.2 LW-Shrinkage for arbitrary covariance structures

In the previous section, it was illustrated that (LW2) is not appropriate for many real world data sets. Hence, I replace (LW2) with a weaker assumption which does not restrict the covariance structure, and introduce a parametrized restriction of the covariance structure. There is a dependency between the limit behavior of LW-Shrinkage and this parameter, hence generalized asymptotics as defined in (LW1) is not used. Instead, shrinkage is analyzed in the LDL and FOLDL, as defined in Section 1.4.3.

In the LDL and FOLDL the sequence of statistical models is indexed by \( p \). Hence, \( \mathbf{X}_p \) denotes a \( p \times n \) matrix of \( n \) observations of \( p \) variables with mean zero and covariance matrix \( \mathbf{C}_p \). \( \mathbf{Y}_p = \mathbf{R}_p^\top \mathbf{X}_p \) denotes the same observations rotated in their eigenbasis, having diagonal covariance \( \mathbf{\Gamma}_p = \mathbf{R}_p^\top \mathbf{C}_p \mathbf{R}_p \). Lower case
letters $x_p^t$ and $y_p^t$ denote the entries of $X_p$ and $Y_p$, respectively.\footnote{The sequence index $p$ and the observation index $t$ are often dropped to improve readability of formulas.}

Making additional weak assumptions, I prove consistency, and, in addition, derive the limit behavior of the shrinkage intensity.

**Modification of (LW2) and (LW3)** I replace (LW2) by a weaker assumption on the moments in the basis of the observations $X$. This does not impose any constraints on the covariance structure:

\[(A2)\quad \exists K_2 : \frac{1}{p_n} \sum_{i=1}^{p_n} \mathbb{E}[(x_{i1}^p)^8] \leq K_2.\]

In the LDL and FOLDL sequences of models indexed by $p$ are considered. Hence, (LW3) has to be modified in the following way:

\[(A3)\quad \frac{\sum_{i,j,k,l \in Q_p} \text{Cov} \left( y_{ij1}^p, y_{k1l}^p \right)^2}{|Q_p|} = o(1).\]

**Assumption on the structure of covariance matrices** (A2) and (A3) allow for a wide range of dispersion and eigenvalue configurations. To further investigate the role of the covariance structure, sequences are categorized by adding an additional parameter $k$. This proves essential for the limit behavior of the optimal shrinkage intensity and for the consistency of LW-Shrinkage. $k$ parameterizes the limit behavior of the normalized dispersion:

\[(A4)\quad \frac{1}{p} \sum_{i} \left( \frac{1}{p} \sum_{j} \gamma_j - \frac{1}{p} \sum_{j} \gamma_j \right)^2 = \Theta \left( \max \left( 1, p^{2k-1} \right) \right).\]

In sequences of models with $k \leq 0.5$ the normalized dispersion is bounded from above and below, as in model A in the last section. For $k > 0.5$ the normalized dispersion grows with the dimensionality, for $k = 1$ it is linear in $p$, as in model B. The case $k = 1$ means that there is no restriction on the covariance structure, larger values of $k$ have no effect.

**Technical assumptions to rule out degenerate cases** The theoretical results in the following sections require two additional technical assumptions: first, it is required that, on average, additional dimensions make a positive contribution to the mean variance: there exists a constant $K_3$ such that

\[(A5)\quad \frac{1}{p} \sum_{i=1}^{p} \mathbb{E} \left[ (x_{i1}^p)^2 \right] \geq K_3.\]

Second, it is required that limits on the relation between second, fourth and eighth moments exist:

\[(A6)\quad \exists \alpha_4, \beta_4 : \quad (1 + \beta_4) \mathbb{E}^2[y_i^2] \leq \mathbb{E}[y_i^4] \leq (1 + \alpha_4) \mathbb{E}^2[y_i^2] ;
(A7)\quad \exists \alpha_8, \beta_8 : \quad (1 + \beta_8) \mathbb{E}^2[y_i^4] \leq \mathbb{E}[y_i^8] \leq (1 + \alpha_8) \mathbb{E}^2[y_i^4].\]

### 3.2.3 LW-Shrinkage in the LDL

The following theorem relates the parameter $k$ to the limit behavior of the optimal shrinkage intensity $\lambda^*$:

**Theorem 3.4** (LDL behavior of LW-Shrinkage). Let (A2), (A4), (A5), (A6) and (A7) hold. Then, there exist $0 < b_l < b_u < 1$

\[
\begin{align*}
    k \leq 0.5 & \quad \Rightarrow \quad \forall p : b_l \leq \lambda^* \leq b_u \\
    k > 0.5 & \quad \Rightarrow \quad \lim_{p \to \infty} \lambda^* = 0.
\end{align*}
\]
3.2. SHRINKAGE FOR ARBITRARY COVARIANCE STRUCTURES

Proof. See Appendix B.4.

The theorem shows that there is a fundamental issue with LW-Shrinkage: if \( k \) is larger than 0.5—all data sets in the last section had \( k = 1 \)—there is no shrinkage in the limit.

This result is surprising as in the literature it is stated that for \( n \) and \( p \) going to infinity at the same rate, the shrinkage intensity converges to a constant, which is in general larger than zero (Ledoit & Wolf, 2004, p. 375). From Theorem 3.4 we learn that this is only true under strong assumptions regarding the structure of the covariance matrix\(^4\). This can be explained in terms of the bias-variance trade-off:

For a general covariance matrix, the standard target has \( p^2 \) biased entries. The sample covariance matrix has \( p^2 \) entries as well, each having higher variance than the target. But when the sample size goes to infinity, the variance of the entries converges to zero while the bias remains unchanged. Therefore, for a general covariance matrix, no shrinkage to the standard target takes place in the LDL.

LDL consistency for arbitrary structures Using Theorem 3.4, it is possible to derive a novel theorem which shows that under the wider assumptions (A2)-(A7), shrinkage remains consistent. As we have seen above, \( \lambda^* \) goes to zero in the LDL for \( k > 0.5 \). In these cases, every estimator of \( \lambda^* \) converging to zero would fulfill \( \mathbb{E}[(\hat{\lambda} - \lambda^*)^2] = o(1) \). Therefore, by considering a relative error, the following theorem shows that the error in the estimator goes to zero faster than the estimator itself:

**Theorem 3.5** (LDL consistency of LW-Shrinkage). Let (A2), (A3), (A4), (A5), (A6) and (A7) hold and

\[
m^p := \mathbb{E} \left[ \left( \frac{\lambda^{*p} - \hat{\lambda}^p}{\lambda^{*p}} \right)^2 \right]
\]

denote the expected squared relative error of the estimate \( \hat{\lambda} \). Then in the LDL holds, independently of \( k \),

\[
m^p = o(1).
\]

Proof. See Appendix B.5.

3.2.4 LW-Shrinkage in the FOLDL

In the previous sections we have seen that values of \( k \) larger than 0.5—or, in other words, fast growing largest eigenvalues—cause a problem in the LDL: although shrinkage remains consistent, the shrinkage intensity converges to zero and no regularization takes place in the limit.

In this section, the FOLDL is considered as an alternative asymptotic framework: \( n \) is kept fixed and only \( p \) tends to infinity. The FOLDL makes covariance estimation more difficult, therefore the low-variance high-bias target should be favored and the shrinkage intensity be prevented from converging to zero. The following theorem quantifies this intuition:

**Theorem 3.6** (FOLDL behavior of LW-Shrinkage). Let (A2), (A4), (A5), (A6) and (A7) hold. Then, there exist \( b_1, b_u \in \mathbb{R}, 0 < b_1 < b_u < 1 \) such that

\[
k < 1 \quad \Rightarrow \quad \lim_{p \to \infty} \lambda^{*p} = 1,
\]

\[
k = 1 \quad \Rightarrow \quad \forall p : b_1 \leq \lambda^{*p} \leq b_u.
\]

Proof. See Appendix B.6.

---

\(^4\)This statement does not hold in general even for \( k \leq 0.5 \): the proof shows that the the sequence of the \( \hat{\lambda}^p \) converges to the \( \lambda^{*p} \), but as there is no relationship between the models in the assumed sequence of statistical models, \( \lambda^{*p} \) does in general not converge.
We see from Theorem 3.6 that in the FOLDL, for all $k < 1$, the shrinkage covariance matrix converges to the target matrix in the limit. Again, one can intuitively understand this in terms of the bias-variance trade-off:

The variance of the sample covariance matrix grows with $p^2$, as the matrix has $p^2$ high-variance entries. For the target, one should ideally look at the bias in the eigenbasis: for large $p$, the bias is approximately equal to the eigenvalue dispersion (see (A4)) and therefore $\Theta(p, p^{2k})$: only for $k = 1$ the error in the target grows as fast as the error in the sample covariance matrix.

In the LDL, consistency results from two averaging effects: on the one hand averaging over an increasing number of samples $n_p$, on the other hand averaging over an increasing number of dimensions $p$.

In the FOLDL, consistency can only result from averaging over an increasing number of dimensions, which puts stronger limits on tolerable dependencies between the $y_i$.

Uncorrelatedness of higher moments  The following extension of (A3) is necessary for consistency of LW-Shrinkage in the FOLDL:

$$
\sum_{i,j,k,l \in Q_p} \text{Cov} \left[ (y_{1i}^p y_{1j}^p)^2, (y_{k1}^p y_{l1}^p)^2 \right] / |Q_p| = o(1).
$$

As in the last section, a relative error is introduced: for $k < 1$, it has to be shown that the error in the estimate goes to zero faster than $\lambda^*$ goes to one, therefore, an error relative to the quantity $1 - \lambda^*$ is of interest here:

**Theorem 3.7** (FOLDL consistency of LW-Shrinkage). Let (A2), (A3), (A3'), (A4), (A5), (A6) and (A7) hold. Let

$$
m^p := \mathbb{E} \left[ \left( \frac{\lambda^* - \hat{\lambda}^p}{1 - \lambda^*} \right)^2 \right].
$$

denote the expected squared relative error of the estimate $\hat{\lambda}$. Then in the FOLDL holds, for $k < 1$,

$$
m^p = o(1).
$$

For $k = 1$, sequences of models exist such that

$$
\exists b_t > 0 : \forall p : m \geq b_t.
$$

In particular, this holds if the $y_i$ are independent.

**Proof.** See Appendix B.7.

The analysis of the consistency of LW-Shrinkage in the FOLDL in Theorem 3.7 provides two important insights: (i) if $k$ is restricted to be smaller than one—as assumption (LW2) does implicitly—infinitely many observations $n$ are not necessary for $\hat{\lambda}$ to converge, instead, the restriction on the dependencies between higher moments in (A3') is sufficient. (ii) Large values of $k$ not only cause weak shrinkage, they also lead to an increased variance of the estimator $\hat{\lambda}$: eigendirections with high variance may dominate the other directions and thereby decrease the advantage gained by averaging over the dimensions. For $k = 1$, this makes LW-Shrinkage inconsistent in the FOLDL.

### 3.2.5 Structural adaptation: automatic orthogonal complement Shrinkage

In the previous sections, it was demonstrated that a large eigenvalue dispersion (corresponding to large values of the parameter $k$) diminishes the advantages of LW-Shrinkage:
3.2. SHRINKAGE FOR ARBITRARY COVARIANCE STRUCTURES

Table 3.1. Results on consistency and limit behavior.

<table>
<thead>
<tr>
<th>Method</th>
<th>$\lambda^{*p} \to 0$</th>
<th>$b_t \leq \lambda^{*p} \leq b_u$</th>
<th>$\lambda^{*p} \to 1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>FOLDL</td>
<td>$p \to \infty, n$ fixed</td>
<td>$k = 1$</td>
<td>consistent</td>
</tr>
<tr>
<td>LDL</td>
<td>$p \to \infty, p/n_p \to c$</td>
<td>$k &gt; 0.5$</td>
<td>consistent</td>
</tr>
<tr>
<td>LDL</td>
<td>$p \to \infty, p/n_p \to c$</td>
<td>$k \leq 0.5$</td>
<td>consistent</td>
</tr>
</tbody>
</table>

- In the LDL, the bias of the shrinkage target becomes higher and higher compared to the variance of the sample covariance matrix. As a consequence, the optimal shrinkage intensity decreases and there is less and less relative improvement in ESE. Mathematically speaking, for $k > 0.5$, $\lambda^*$ goes to zero. Only for $k < 0.5$, $\lambda^*$ is bounded away from zero.

- In the FOLDL, the shrinkage target remains useful, but the estimator of the shrinkage intensity, $\hat{\lambda}$, suffers from large variance. Mathematically speaking, $\lambda^*$ stays bounded away from zero for $k = 1$, but the estimator $\hat{\lambda}$ is no longer consistent.

Table 3.1 summarizes the results from the previous sections.

Shrinkage on the orthogonal complement To obtain a finite shrinkage intensity for large $p$ and $n$, I propose an extension of LW-Shrinkage called oc-Shrinkage: it restricts shrinkage to the orthogonal complement $(oc)$ of the first eigendirection.

Figure 3.3 illustrates this approach: it shows a three-dimensional population covariance matrix (green) with a high dispersion that makes it highly ellipsoidal. The result is a high level of discrepancy between the spherical standard shrinkage target (blue) and the population covariance. The convex combination of target and sample covariance (red) closest to the population covariance puts extremely low weight on the target.

In the orthogonal complement of the first eigendirection of the sample covariance matrix, the standard target calculated on the whole space is far too large (dashed sphere). For the target calculated only on the complement, the discrepancy between sample covariance and target is strongly reduced: stronger shrinkage takes place and higher error reduction is achieved.

Assuming a single growing eigenvalue To simplify the theoretical analysis, the case with a single growing eigenvalue, while the remaining eigenvalues stay bounded, is considered:

$$\eta_1 := p^{-k/2} y_1,$$
$$\eta_i := y_i \quad 2 \leq i \leq p,$$

there exist constants $F_l$ and $F_u$ such that

$$\forall i: F_l \leq \mathbb{E}[\eta_i^2] \leq F_u.$$

(A4')

Convergence of the eigendecomposition In order to prove that the projection onto the empirical orthogonal complement $(\hat{oc})$ does not affect the consistency of the estimator $\hat{\lambda}_{\hat{oc}}$, the error in the estimation of the first eigendirection has to be quantified. I derive a generalization of Theorem 2.3 in (Nadler, 2008) which is valid without assuming Gaussianity of the noise and existence of fourth moment of the signal:

**Theorem 3.8** (convergence of the eigendecomposition). Consider a spiked covariance model

$$x = u\eta_1 + \sigma e,$$
where \( u \) and \( \epsilon \) are zero mean, unit variance. Then, in the LDL,

\[
\gamma_1 = \begin{cases} 
\sigma^2 \left(1 + \frac{\sigma^2}{p} \right)^2, & \text{if } n/p < \sigma^4/\|v_1\|^4, \\
\left(\|v_1\|^2 + \sigma^2\right) \left[1 + \frac{p}{n} \frac{\sigma^2}{\|v_1\|^2}\right], & \text{if } n/p \geq \sigma^4/\|v_1\|^4,
\end{cases}
\]  

(3.8)

and for the dot product between population and sample eigenvector we have

\[
R^2(p/n) := \|\langle \tilde{v}_1, v_i \rangle\| = \begin{cases} 
0 & \text{if } n/p < \sigma^4/\|v_1\|^4, \\
\frac{n/p \cdot \|v_1\|^4/\sigma^4 - 1}{n/p \cdot \|v_1\|^4/\sigma^4 + \sigma^4/\|v_1\|^4} & \text{if } n/p \geq \sigma^4/\|v_1\|^4.
\end{cases}
\]  

(3.9)

Proof. See Appendix B.8.

With Theorem 3.8, eq. (3.9), it is possible to prove that oc-Shrinkage is consistent.

**Consistency of oc-Shrinkage** For the theoretical analysis, a new orthonormal set of vectors \( \tilde{v}_i \) is defined, based on the sample and population eigenvectors \( \tilde{v}_i \) and \( v_i \):

\[
\tilde{v}_1 = v_1 \quad \text{and} \quad \langle \tilde{v}_i, v_j \rangle = 0 \quad \text{if} \quad j \neq i, j \neq 1.
\]  

(3.10)

For the rotated data set \( \mathbf{U} = \mathbf{VY} \), we then have

\[
u_{it} = y_{it}(\tilde{v}_i, v_i) + y_{it}(\tilde{v}_i, v_1) := y_{it}\sqrt{1 - \varepsilon_i^2} + y_{1t}\tilde{\varepsilon}_i,
\]  

(3.11)

making the population covariance matrix \( \mathbf{C}^{oc} \) an arrowhead matrix (Nadler, 2008). The \( \varepsilon_i \) quantify the error in the estimate of the first eigendirection: if all \( \varepsilon_i \) are zero, the first sample and population
eigendirection coincide. With this basis, we have
\[
\hat{\lambda}_{\hat{b}oc} := \frac{\sum_{i,j \geq 2} \left\{ \text{Var}\left( \hat{S}_{ij}^c \right) - \text{Cov}\left( \hat{S}_{ij}^c, \hat{\nu}_{\hat{b}oc} I_{ij} \right) \right\}}{\sum_{i,j \geq 2} \left( \hat{S}_{ij}^c - \hat{\nu}_{\hat{b}oc} I_{ij} \right)^2},
\]
with \( \hat{\nu}_{\hat{b}oc} = (p - 1)^{-1} \sum_{i,j \geq 2} \hat{S}_{ij}^c \).

As the orthogonal complement is an empirical quantity, we have to consider the squared error (SE) of
the convex combination (compare to eq. (3.3))
\[
\hat{\Delta}_{\hat{b}oc}(\lambda) := \sum_{i,j \geq 2} \left| C_{ij}^c - (1 - \lambda) \hat{S}_{ij}^c - \lambda \hat{\nu}_{\hat{b}oc} I_{ij} \right|^2,
\]
the optimal shrinkage intensity is then given by
\[
\lambda_{\hat{b}oc} := \arg \min_{\lambda} \hat{\Delta}_{\hat{b}oc}(\lambda).
\]

**Theorem 3.9** (LDL consistency of oc-Shrinkage). Let (A2), (A3), (A4'), (A5), (A6) and (A7) hold. In
addition, assume that the 16\(^{th}\) moments of the \( y_i \) exist and are bounded. Then, independently of \( k \), in the
LDL holds
\[
\left( \hat{\lambda}_{\hat{b}oc} - \lambda_{\hat{b}oc}^* \right)^2 = o(1).
\]

**Proof.** See Appendix B.9. \( \square \)

Here, the existence of 16\(^{th}\) moments assures that fourth moments exist for all fourth order terms in
the LW-Shrinkage formula. This facilitates the proof of consistency, as it allows first taking the limit and
then the maximum over dimension (see the proof in Appendix B.9).

**Correcting the largest eigenvalue** In addition, Theorem 3.8 yields an improvement over just keeping
the sample eigenvalue: eq. (3.8) allows for a correction of the largest eigenvalue:
\[
\hat{\gamma}_1 = \gamma_1 + \frac{p \sigma^2}{n} \frac{\gamma_1}{\gamma_1 - \sigma^2} \approx \gamma_1 + \frac{p \sigma^2}{n}
\]
\[
\iff \gamma_1 \approx \hat{\gamma}_1 - \frac{p \sigma^2}{n} \approx \sum_{i=2}^p \hat{\gamma}_i - \frac{1}{n} \sum_{i=2}^p \hat{\gamma}_i.
\]

**Automatic model selection** In the previous section, it was shown that oc-Shrinkage consistently
estimates the optimal shrinkage intensity on the orthogonal complement and is therefore better or at least
as good as LW-Shrinkage. For finite sample sizes, this is not guaranteed. Hence one has to decide whether
applying oc-Shrinkage is likely to yield an improvement in squared error (SE).

For this, the improvement on the sample orthogonal complement in terms of ESE and SE is analyzed:
\[
\delta_{\hat{b}oc}(\hat{\lambda}, \hat{\nu}, \hat{\lambda}_{\hat{b}oc}, \hat{\nu}_{\hat{b}oc}) := \frac{1}{p} \left( \Delta_{\hat{b}oc}(\hat{\lambda}, \hat{\nu}) - \Delta_{\hat{b}oc}(\hat{\lambda}_{\hat{b}oc}, \hat{\nu}_{\hat{b}oc}) \right),
\]
\[
\delta_{\hat{b}oc}(\hat{\lambda}, \hat{\nu}, \hat{\lambda}_{\hat{b}oc}, \hat{\nu}_{\hat{b}oc}) := \frac{1}{p} \left( \Delta_{\hat{b}oc}(\hat{\lambda}, \hat{\nu}) - \Delta_{\hat{b}oc}(\hat{\lambda}_{\hat{b}oc}, \hat{\nu}_{\hat{b}oc}) \right).
\]

It is now possible to directly estimate the performance difference in ESE on the sample orthogonal complement using eq. (3.4) in the sample eigenbasis:
\[
\delta_{\hat{b}oc}(\hat{\lambda}, \hat{\nu}, \hat{\lambda}_{\hat{b}oc}, \hat{\nu}_{\hat{b}oc}) = \sum_{i \geq 2, j \geq 2} \frac{1}{p} \left\{ -2 \hat{\lambda} \left( \text{Var}\left( \hat{S}_{ij}^c \right) - \text{Cov}\left( \hat{S}_{ij}^c, \hat{\nu} I_{ij} \right) \right) \right\}_{i \neq j, p \neq \hat{\nu}}.
\]
Algorithm 2: aoc-Shrinkage

Input: data $X$, sample eigenvalues $\hat{\gamma}$, sample eigenvectors $\hat{V}$
Output: aoc-Shrinkage covariance matrix $C_{\text{aoc-shr}}$

1. $r^* = 0$, $r = 1$
2. $\lambda_{\text{old}} = \text{calc}_\lambda(X)$
3. $\nu_{\text{old}} = \text{mean}(\hat{\gamma})$
4. while TRUE
5. $X_{\text{oc}} = \hat{V}(:, r + 1 : p)^T X$
6. $\hat{\gamma}_{\text{oc}} = \hat{\gamma}(r + 1 : p)$
7. $\lambda_{\text{new}} = \text{calc}_\lambda(X_{\text{oc}})$
8. $\nu_{\text{new}} = \text{mean}(\hat{\gamma}_{\text{oc}})$
9. $\delta = \text{calc}_\Delta(X_{\text{oc}}, \lambda_{\text{new}}, \nu_{\text{new}}, \lambda_{\text{old}}, \nu_{\text{old}})$
10. $\sigma_{\Delta} = \text{calc}_\sigma_Delta(\hat{\gamma}_{\text{oc}}, \lambda_{\text{new}}, \nu_{\text{new}}, \lambda_{\text{old}}, \nu_{\text{old}})$
11. $\sigma_{\text{E}} = \text{calc}_\sigma_E(\hat{\gamma}_{\text{oc}})$
12. if $\delta - m_{\Delta}\sigma_{\Delta} - m_{\text{E, oc}}\lambda_{\text{new}}^2 \sigma_{\text{E}} > 0$
13. $r^* = r^* + 1$
14. $r = r + 1$
15. $\lambda_{\text{old}} = \lambda_{\text{new}}$
16. $\nu_{\text{old}} = \nu_{\text{new}}$
17. else
18. BREAK
19. end
20. $\hat{P}_{\text{pca}} = \hat{V}(:, 1 : r^*)^T \hat{V}(:, 1 : r^*)$
21. $\hat{P}_{\text{oc}} = \hat{V}(:, r^* + 1 : p)^T \hat{V}(:, r^* + 1 : p)$
22. $C_{\text{shr}} = (1 - \lambda_{\text{old}}) \cdot S + \hat{\lambda}_{\text{old}} \hat{\nu}_{\text{old}} \text{eye}(p)$
23. return $C_{\text{aoc-shr}} = \hat{P}_{\text{pca}}^T S \hat{P}_{\text{pca}} + \hat{P}_{\text{oc}}^T C_{\text{shr}} \hat{P}_{\text{oc}}$

where $\hat{\lambda}$ is the shrinkage intensity estimated on the whole space and $\hat{\lambda}_{\text{oc}}$ is estimated on the orthogonal complement of the first sample eigendirection.

Asymptotically, the estimate of the performance difference converges to the performance difference in SE:

**Theorem 3.10** (convergence of the error). Let (A2), (A3), (A4'), (A5), (A6) and (A7) hold. In addition, assume that $16^{th}$ moments of the $y_i$ exist and are bounded. Then, independently of $k$, in the LDL holds

$$\delta_{\text{oc}}(\lambda, \hat{\nu}, \hat{\lambda}_{\text{oc}}, \hat{\nu}_{\text{oc}}) - \delta_{\text{oc}}(\lambda, \nu, \hat{\lambda}_{\text{oc}}, \hat{\nu}_{\text{oc}}) = o(1).$$

**Proof.** See Appendix B.10.

Nevertheless, always applying oc-Shrinkage when $\delta_{\text{oc}}$ is positive does not yield good results: $\delta_{\text{oc}}$ has a positive bias. There are two main reasons for this which should be taken into account:

$$+ 2\hat{\lambda}_{\text{oc}} \left( \text{Var} \left( \hat{S}^{I_{1j}}_{ij} \right) - \text{Cov} \left( \hat{S}^{I_{1j}}_{ij}, \hat{\nu}_{\text{oc}} I_{ij} \right) \right)$$

$$+ \hat{\lambda}^2 \left( \hat{S}^{I_{1j}}_{ij} - \hat{\nu} I_{ij} \right)^2 - \hat{\lambda}^2_{\text{oc}} \left( \hat{S}^{I_{1j}}_{ij} - \hat{\nu}_{\text{oc}} I_{ij} \right)^2$$

$$=: p \hat{E}_{\text{oc}}$$

$$=: -2\hat{V}_{\text{oc}} + 2\hat{\lambda}_{\text{oc}} \hat{V} + \hat{\lambda}^2 \hat{E} - \hat{\lambda}^2_{\text{oc}} \hat{E}_{\text{oc}},$$
3.2. SHRINKAGE FOR ARBITRARY COVARIANCE STRUCTURES

- \( \hat{\lambda}_{\text{oc}} \) is defined as the minimizer of \( \Delta_{\text{oc}}(\lambda, \hat{\nu}_{\text{oc}}) \), therefore \( \Delta_{\text{oc}}(\hat{\lambda}_{\text{oc}}, \hat{\nu}_{\text{oc}}) \) underestimates the error. Subtracting a suitable multiple \( m_\Delta \) of the standard deviation \( \hat{\sigma}_{\Delta_{\text{oc}}}(\lambda_{\text{oc}}, \hat{\nu}_{\text{oc}}) \) reduces this bias.

- \( \hat{\nu}_{\text{oc}} \) is the average of the eigenvalues excluding the largest one. Given that the largest eigenvalue has a positive bias, \( \hat{\nu}_{\text{oc}} \) has a negative bias. Therefore \( \hat{E}_{\text{oc}} = \sum_{i,j}(S_{ij} - \hat{\nu}_{\text{oc}} I_{ij})^2 \) underestimates the error as well. This underestimation is mainly relevant for small \( p \), when \( \hat{\nu} \) is an average over only a few dimensions. Adding a suitable multiple \( m_\nu \) of the standard deviation \( \sigma_{\hat{E}_{\text{oc}}} \) to \( \hat{E}_{\text{oc}} \) reduces this bias.

An approximation of the standard deviation \( \hat{\sigma}_{\Delta_{\text{oc}}}(\lambda_{\text{oc}}, \hat{\nu}_{\text{oc}}) \) is given by

\[
\hat{\sigma}^2_{\Delta_{\text{oc}}}(\lambda_{\text{oc}}, \hat{\nu}_{\text{oc}}) = \text{Var} \left( -2\hat{\lambda}_{\text{oc}} \hat{V}_{\text{oc}} + 2\lambda_{\text{oc}} \hat{V} + \hat{\lambda}^2 \hat{E} - \hat{\lambda}^2 \hat{E}_{\text{oc}} \right)
\approx \text{Var} \left( -2(\hat{\lambda} - \lambda_{\text{oc}}) \hat{V}_{\text{oc}} + (\hat{\lambda}^2 - \hat{\lambda}_{\text{oc}}^2) \hat{E}_{\text{oc}} \right)
= 4(\hat{\lambda} - \lambda_{\text{oc}})^2 \sigma^2_{\hat{V}_{\text{oc}}} + (\hat{\lambda}^2 - \hat{\lambda}_{\text{oc}}^2)^2 \sigma^2_{\hat{E}_{\text{oc}}}
- 4(\hat{\lambda} - \lambda_{\text{oc}})(\hat{\lambda}^2 - \hat{\lambda}_{\text{oc}}^2) \sigma_{\hat{V}_{\text{oc}}} \sigma_{\hat{E}_{\text{oc}}} \text{corr}(\hat{V}_{\text{oc}}, \hat{E}_{\text{oc}}).
\]

A parametric bootstrap assuming normality and covariance \( \hat{S} \) on the orthogonal complement yields estimates \( \hat{\sigma}_{\hat{E}_{\text{oc}}}, \hat{\sigma}_{\hat{V}_{\text{oc}}} \), and \( \text{corr}(\hat{V}_{\text{oc}}, \hat{E}_{\text{oc}}) \). Closed form solutions for the results of the parametric bootstrap are provided in Appendix B.11.

The exact values of multipliers \( m_\Delta \) and \( m_\nu \) do not have a strong influence on the performance. Setting \( m_\Delta = 0.45 \), which corresponds to 75\% probability of improvement under Gaussianity, and \( m_\nu = 0.1 \), seems to be sufficient. This yields the conservative estimate

\[
\hat{\delta}_{\text{oc}}^\text{cons}(\hat{\lambda}, \hat{\nu}, \hat{\lambda}_{\text{oc}}, \hat{\nu}_{\text{oc}}) = \hat{\delta}_{\text{oc}}(\hat{\lambda}, \hat{\nu}, \hat{\lambda}_{\text{oc}}, \hat{\nu}_{\text{oc}}) - m_\Delta \hat{\sigma}_{\Delta_{\text{oc}}}(\lambda_{\text{oc}}, \hat{\nu}_{\text{oc}}) - m_\nu \hat{\lambda}_{\text{oc}}^2 \hat{\sigma}_{\hat{E}_{\text{oc}}}.
\]

A data set may have multiple large eigenvalues. It is straightforward to iterate the procedure and thus automatically select the number of retained eigendirections \( \hat{r} \). In the following, this is called automatic orthogonal complement Shrinkage (aoc-Shrinkage, see Algorithm 2).

The computational cost of aoc-Shrinkage is larger than that of LW-Shrinkage as it additionally requires an eigendecomposition. In the applications considered here, this additional cost is negligible, as the eigendecomposition can be used for the matrix inversions in LDA, QDA or portfolio optimization.

3.2.6 Simulation results

The proposed method is tested in a set of simulations based on model B from eq. (3.7). Because the population covariance matrix is known, it is possible to measure the percentage improvement in average loss (PRIAL) over the sample covariance matrix of LW-Shrinkage, shrinkage on the orthogonal complement (oc-Shrinkage) and the automatic selection of the optimal orthogonal complement size (aoc-Shrinkage).

Simulation 1 In this simulation, the dimensionality \( p \) is varied in model B and each of the methods is applied (\( P_1 = 0.5 \)). Figure 3.4 reports the PRIAL of LW-Shrinkage, oc-Shrinkage for one eigendirection and aoc-Shrinkage.

The asymptotic behavior of LW-Shrinkage described in Theorem 3.4 is obvious: in the LDL, the shrinkage intensity and, as a consequence, the improvement over the sample covariance matrix go to zero. In contrast, aoc-Shrinkage achieves constant PRIAL at a reduced variance.

For small dimensions, and therefore small largest eigenvalues, LW-Shrinkage outperforms oc-Shrinkage. There are two reasons for this: (i) the largest sample eigenvalue is spurious, its retention increases the MSE compared to shrinking it as in LW-Shrinkage. (ii) By performing shrinkage on the orthogonal complement, the precision of the estimator for target and shrinkage intensity is decreased.
The automatic procedure aoc-Shrinkage applies the better method in each regime: For small $p$ it chooses LW-Shrinkage, while for large $p$ it chooses oc-Shrinkage. The procedure is conservative: it only switches to oc-Shrinkage if there is a high probability that this leads to an improvement in the MSE.

**Simulation 2** In this simulation, the effect of the size of the largest eigenvalue for constant dimensionality and number of observations is analyzed, governed by the parameter $P_1$ in model B.

Figure 3.5 shows that while for high values of $P_1$ (dominating largest eigenvalues) oc-Shrinkage outperforms LW-Shrinkage, it does not achieve constant PRIAL: with increasing largest eigenvalue, at some
3.2. SHRINKAGE FOR ARBITRARY COVARIANCE STRUCTURES

Figure 3.6. PRIAL vs. correlation probability on the orthogonal complement in Simulation 2. Dashed Lines show one standard deviation. $n = 600, p = 300$. Average over $K = 100$ runs.

Figure 3.7. Condition number in Simulation 2. $n = 600, p = 300$. Average over $K = 100$ runs.

point the error in the corresponding direction dominates the error in the orthogonal complement.

Figure 3.6 shows that in the population orthogonal complement, the PRIAL of oc-Shrinkage is constant, while for LW-Shrinkage it still tends to zero. As a consequence, the condition number of the LW-Shrinkage covariance matrix approaches the one of the sample covariance matrix while oc-Shrinkage covariance matrices remain better conditioned than the population covariance matrix (Figure 3.7).

The automatic procedure aoc-Shrinkage performs optimally for small and dominating largest eigen-values (small and large $P_1$).
Simulation 3  To test the iterative approach for the case of multiple large eigenvalues in \( \text{aoc-Shrinkage} \), model B from eq. (3.7) is extended to three signals, \( P_i = (0.1, 0.25, 0.5) \). Figure 3.8 shows that LW-Shrinkage again behaves as predicted by Theorem 3.4: the shrinkage intensity and, as a consequence, the PRIAL converge to zero in the LDL. The same holds for oc(1) and oc(2), orders of oc-Shrinkage lower than the number of signals, although performance degrades more slowly.

For small dimensionalities eigenvalues are small and therefore there is no advantage for oc-Shrinkage. On the contrary, the higher the order of oc-Shrinkage, the larger the error by projecting out spurious large eigenvalues which should have been subject to regularization. The automatic order selection \( \text{aoc-Shrinkage} \) yields nearly optimal PRIAL for all dimensionalities.

Table 3.2. Portfolio risk. Mean absolute deviations \( 10^3 \) (mean squared deviations \( 10^6 \)) of the resulting portfolios for the different covariance estimators and markets. \( \dagger \) := aoc-Shrinkage significantly better than this model at the 5% level, tested by a randomization test.

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>EU</th>
<th>HK</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample covariance</td>
<td>8.56(\dagger) (156.1(\dagger))</td>
<td>5.93(\dagger) (78.9(\dagger))</td>
<td>6.57(\dagger) (81.2(\dagger))</td>
</tr>
<tr>
<td>LW-Shrinkage</td>
<td>6.27(\dagger) (86.4(\dagger))</td>
<td>4.43(\dagger) (46.2(\dagger))</td>
<td>6.32(\dagger) (76.2(\dagger))</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>0.09</td>
<td>0.12</td>
<td>0.10</td>
</tr>
<tr>
<td>LW-FMT-Shrinkage</td>
<td>5.56(\dagger) (69.6(\dagger))</td>
<td>4.00(\dagger) (39.1(\dagger))</td>
<td>6.17(\dagger) (72.9(\dagger))</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>0.41</td>
<td>0.44</td>
<td>0.42</td>
</tr>
<tr>
<td>aoc-Shrinkage</td>
<td>5.41 (67.0)</td>
<td>3.83 (36.3)</td>
<td>6.11 (71.8)</td>
</tr>
<tr>
<td>( \lambda )</td>
<td>0.75</td>
<td>0.79</td>
<td>0.75</td>
</tr>
<tr>
<td>average ( \hat{\rho} )</td>
<td>1.64</td>
<td>1.17</td>
<td>1.41</td>
</tr>
</tbody>
</table>

3.2.7  Real world data results

Real world data I: portfolio optimization In finance, accurate covariance estimates are needed for risk minimization. In the unregularized portfolio optimization setting described in section 2.3, aoc-
3.2. SHRINKAGE FOR ARBITRARY COVARIANCE STRUCTURES

Shrinkage is compared to the sample covariance, LW-Shrinkage and LW-FMT-Shrinkage (factor model shrinkage target).

Results in Table 3.2 show that shrinkage-based approaches clearly outperform the sample covariance matrix. Despite the unfavorable ratio of observations to dimensionality, LW-Shrinkage has very low values of $\lambda$: the stocks are highly correlated and therefore the standard target is highly inappropriate. The factor model in LW-FMT-Shrinkage is a more suitable target; it leads to stronger shrinkage and better portfolios. The proposed aoc-Shrinkage yields even stronger shrinkage and significantly outperforms all of the remaining tested methods. The automatic approach excludes one to two dimensions from shrinkage.

Comparing to the results of Factor Analysis in Table 3.2, it can be seen that aoc-Shrinkage yields a performance comparable to DVA Factor Analysis.

Table 3.3. aoc-Shrinkage accuracies for classification tasks on ISOLET and USPS data. * := significantly better than all other methods at the 5% level, tested by a randomization test.

<table>
<thead>
<tr>
<th></th>
<th>ISOLET 500</th>
<th>ISOLET 2000</th>
<th>ISOLET 5000</th>
<th>USPS 500</th>
<th>USPS 2000</th>
<th>USPS 5000</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard LDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>accuracy</td>
<td>75.77%</td>
<td>92.29%</td>
<td>94.1%</td>
<td>72.31%</td>
<td>87.45%</td>
<td>89.56%</td>
</tr>
<tr>
<td>LW-Shrinkage-LDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>accuracy</td>
<td>88.92%</td>
<td>93.25%</td>
<td>94.3%</td>
<td>83.77%</td>
<td>88.37%</td>
<td>89.77%</td>
</tr>
<tr>
<td>average $\hat{\lambda}$</td>
<td>0.082</td>
<td>0.022</td>
<td>0.0089</td>
<td>0.081</td>
<td>0.021</td>
<td>0.0087</td>
</tr>
<tr>
<td>aoc-Shrinkage-LDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>accuracy</td>
<td>89.69%*</td>
<td>93.42%*</td>
<td>94.33%*</td>
<td>83.95%*</td>
<td>88.37%</td>
<td>89.77%</td>
</tr>
<tr>
<td>average $\hat{\lambda}$</td>
<td>0.17</td>
<td>0.041</td>
<td>0.013</td>
<td>0.097</td>
<td>0.021</td>
<td>0.0087</td>
</tr>
<tr>
<td>average $\hat{\nu}$</td>
<td>3.21</td>
<td>2.06</td>
<td>1</td>
<td>1.23</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>standard QDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>accuracy</td>
<td>2.78%</td>
<td>4.88%</td>
<td>14.09%</td>
<td>10.11%</td>
<td>49.45%</td>
<td>72.43%</td>
</tr>
<tr>
<td>LW-Shrinkage-QDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>accuracy</td>
<td>58.57%</td>
<td>75.4%</td>
<td>79.25%</td>
<td>82.2%</td>
<td>88.85%</td>
<td>89.67%</td>
</tr>
<tr>
<td>average $\hat{\lambda}$</td>
<td>0.53</td>
<td>0.23</td>
<td>0.11</td>
<td>0.28</td>
<td>0.091</td>
<td>0.039</td>
</tr>
<tr>
<td>aoc-Shrinkage-QDA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>accuracy</td>
<td>59.51%</td>
<td>80.84%</td>
<td>87.35%</td>
<td>83.31%</td>
<td>89.4%*</td>
<td>90.07%</td>
</tr>
<tr>
<td>average $\hat{\lambda}$</td>
<td>0.57</td>
<td>0.37</td>
<td>0.24</td>
<td>0.35</td>
<td>0.16</td>
<td>0.067</td>
</tr>
<tr>
<td>average $\hat{\nu}$</td>
<td>0.1373</td>
<td>1.1727</td>
<td>2.6169</td>
<td>0.673</td>
<td>2.132</td>
<td>1.949</td>
</tr>
</tbody>
</table>

Real world data II: USPS and ISOLET  Linear and Quadratic Discriminant Analysis (LDA and QDA, (Hastie et al., 2008)) are examples for widely used algorithms which rely on precise covariance estimates. Both estimate Gaussian distributions for each class and assign a new data point to the class which yields the highest likelihood. While LDA assumes that all classes share a common covariance matrix and only differ in the mean, QDA estimates the covariance matrix of each class separately. Standard LDA and QDA are based on the sample covariance matrix.

To evaluate the covariance estimators, LDA and QDA are applied to hand-written digit recognition (USPS) and spoken letter recognition (ISOLET), both described in section 1.6.

Table 3.3 reports classification accuracies on USPS and ISOLET for three training set sizes, averaged over 100 random splits into training and test sets. The proposed aoc-Shrinkage outperforms LW-Shrinkage for QDA and LDA in nearly all cases. Only for LDA and large sample sizes on the relatively low-dimensional USPS data, there is no difference between LW- and aoc-Shrinkage: the automatic procedure decides that LW-Shrinkage on the whole space is optimal.
Table 3.4. Accuracies for classification tasks on NU-ERP-BCI data. Artificially injected noise in one electrode. * := significantly better than all other methods at the 5% level, tested by a randomization test.

<table>
<thead>
<tr>
<th>$\sigma_{\text{noise}}$</th>
<th>0</th>
<th>10</th>
<th>30</th>
<th>100</th>
<th>300</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>standard LDA</td>
<td>92.28%</td>
<td>92.28%</td>
<td>92.28%</td>
<td>92.28%</td>
<td>92.28%</td>
<td>92.28%</td>
</tr>
<tr>
<td>LW-Shrinkage-LDA</td>
<td>92.39%</td>
<td>92.94%</td>
<td>92.18%</td>
<td>88.04%</td>
<td>82.15%</td>
<td>73.79%</td>
</tr>
<tr>
<td>$\hat{\lambda}$</td>
<td>0.028</td>
<td>0.006</td>
<td>0.0022</td>
<td>0.002</td>
<td>0.002</td>
<td>0.002</td>
</tr>
<tr>
<td>aoc-Shrinkage-LDA</td>
<td>93.27%*</td>
<td>93.27%*</td>
<td>93.24%*</td>
<td>92.88%*</td>
<td>93.16%*</td>
<td>93.19%*</td>
</tr>
<tr>
<td>$\hat{\lambda}$</td>
<td>0.033</td>
<td>0.033</td>
<td>0.033</td>
<td>0.033</td>
<td>0.033</td>
<td>0.033</td>
</tr>
<tr>
<td>average $\hat{r}$</td>
<td>2.0836</td>
<td>3.0945</td>
<td>3.0891</td>
<td>3.0891</td>
<td>3.0891</td>
<td>3.09</td>
</tr>
</tbody>
</table>

Figure 3.9. High-variance components responsible for failure of LW-Shrinkage. The left two components correspond to ocular and probably facial muscle artifacts. The right component detects the channel noise direction artificially injected in the simulation. $\sigma_{\text{noise}} = 10$. Subject 1.

The average number of eigendirections excluded from shrinkage, $\hat{r}$, goes up to 3.21. If covariances are estimated on very few samples (QDA with 500 data points), aoc-Shrinkage cannot reliably estimate the improvement obtained by restricting shrinkage to the orthogonal complement and therefore defaults to LW-Shrinkage. For intermediate sample sizes (QDA with 5000 or LDA with 500 data points) $\hat{r}$ is largest. For large sample sizes (LDA with 5000 data points) the sample covariance is a good estimate, the shrinkage intensity is small and only little improvement is possible.

Real world data III: NU-ERP-BCI LW-Shrinkage has been successfully applied to Linear Discriminant Analysis in a BCI context: LW-Shrinkage-LDA is the state-of-the-art for the classification of event related potentials (Vidaurre et al., 2009, 2011; Lemm et al., 2011; Blankertz et al., 2011).

The noise sensitivity of LW-Shrinkage in LDA in the NU-ERP-BCI, described in section 1.6, is analyzed by injecting Gaussian noise of varying strength in a random electrode ($n_{\text{train}} = 1000$, 100 repetitions). Table 3.4 shows that for all noise levels, aoc-Shrinkage-LDA outperforms standard and LW-Shrinkage-LDA. Without noise, the automatic approach excludes $\hat{r} \approx 2$ high-variance directions from shrinkage. With injected noise, the number of excluded directions increases to $\hat{r} \approx 3$, as the procedure detects the additional high-variance noise component (depicted to the right in Figure 3.9) and adapts shrinkage such that the performance remains unaffected.

For LW-Shrinkage, noise affects the analytic regularization and thereby has an influence on the performance. Interestingly, a small amount of noise improves LW-Shrinkage for this data set by reducing regularization. A high noise level, on the other hand, reduces regularization too much and performance degrades.
3.3 Covariance shrinkage for autocorrelated data

Ledoit-Wolf Shrinkage and most variations, including aoc-Shrinkage, assume i.i.d. data. Real world time series, however, are often non-i.i.d. as they possess pronounced autocorrelation (AC). This makes covariance estimation in high dimensions even harder: the data dependence lowers the effective sample size available for constructing the estimator (Thiébaux & Zwiers, 1984). Thus, stronger regularization is needed.

In Figure 3.10 the simple case of an autoregressive model of lag order 1 (AR(1)-model) serves as an example for an arbitrary generative model with autocorrelation. The figure shows, for three levels of autocorrelation (left), the population and sample eigenvalues (middle): with increasing autocorrelation the sample eigenvalues become more biased. This bias is an optimistic measure for the quality of the covariance estimator: it neglects that population and sample eigenbasis also differ (Ledoit & Wolf, 2012). Comparing sample eigenvalues to the population variance in the sample eigenbasis, the bias is even larger (right).

In practice, violations of the i.i.d. assumption are often ignored (Lotte & Guan, 2011; Samek et al., 2013; Gramfort et al., 2014), although Sancetta (2008) proposed a consistent shrinkage estimator under autocorrelation. In this section, I contribute by showing in theory, simulations and on real world data, that

(i) ignoring autocorrelations in LW-Shrinkage leads to large estimation errors.

(ii) for finite samples Sancetta’s estimator is still substantially biased and highly sensitive to the number of incorporated time lags.

I propose a new bias-corrected estimator which

(iii) outperforms LW-Shrinkage and Sancetta’s method in the presence of autocorrelation.

(iv) is robust to the choice of the lag parameter.

3.3.1 Extension of LW-Shrinkage for autocorrelated data

Let us recapitulate that the analytic shrinkage estimator of Ledoit and Wolf is obtained by replacing expectations with sample estimates:

\[
\text{Var}(\hat{S}_{ij}) \rightarrow \hat{\text{Var}}(\hat{S}_{ij}) := \frac{1}{n^2} \sum_{s=1}^{n} \left( x_{is} x_{js} - \frac{1}{n} \sum_{t=1}^{n} x_{it} x_{jt} \right)^2 \quad (3.12)
\]

\[
\mathbb{E} \left[ (\hat{S}_{ij} - \hat{T}_{ij})^2 \right] \rightarrow \hat{\mathbb{E}} \left[ (\hat{S}_{ij} - \hat{T}_{ij})^2 \right] := (\hat{S}_{ij} - \hat{T}_{ij})^2. \quad (3.13)
\]
While the estimator eq. (3.13) is unbiased even under a violation of the i.i.d. assumption, the estimator eq. (3.12) is based on

$$\text{Var} \left( \frac{1}{n} \sum_{t=1}^{n} x_{it}x_{jt} \right) \overset{\text{i.i.d.}}{=} \frac{1}{n} \text{Var} \left( x_{it}x_{jt} \right).$$

If the data are autocorrelated, cross terms cannot be ignored and we obtain

$$\text{Var} \left( \frac{1}{n} \sum_{t=1}^{n} x_{it}x_{jt} \right) = \frac{1}{n^2} \sum_{s,t=1}^{n} \text{Cov} \left( x_{it}x_{jt}, x_{is}x_{js} \right) = \frac{1}{n} \text{Cov} \left( x_{it}x_{jt}, x_{it}x_{jt} \right) + \frac{2}{n} \sum_{s=1}^{n-1} \frac{n-s}{n} \text{Cov} \left( x_{it}x_{jt}, x_{i,t+s}x_{j,t+s} \right) =: \frac{1}{n} \gamma_{ij}(0) + 2 \sum_{s=1}^{n-1} \gamma_{ij}(s).$$

Figure 3.11 illustrates the effect of ignoring the cross terms for increasing autocorrelation (larger AR-coefficients, see Section 3.3.3 for details on the simulation). It compares LW-Shrinkage to an oracle shrinkage based on the population variance of the sample covariance\(^5\).

With increasing AC, the population variance of \(S\) increases because the effective sample size is reduced (Thiébaux & Zwiers, 1984), yet the LW-Shrinkage variance estimator eq. (3.12) does not increase (outer left). As a consequence, for oracle shrinkage the shrinkage intensity increases. For the LW-Shrinkage estimator it even decreases because the denominator in eq. (3.5) grows (middle left). With increasing autocorrelation, the sample covariance becomes a less precise estimator: for optimal (stronger) shrinkage more improvement becomes possible, yet LW-Shrinkage does not improve (middle right).

Looking at the variance estimates in the sample eigendirections for AR-coefficients of 0.7, we see that the bias of LW-Shrinkage is only marginally smaller than the bias of the sample covariance, while oracle shrinkage yields a substantial bias reduction (outer right).

**Sancetta-Estimator** Sancetta (2008) proposed an estimator for eq. (3.14):

$$\mathcal{Y}_{ij}^{\text{San}}(s) := \frac{1}{n} \sum_{t=1}^{n-s} \left( x_{it}x_{jt} - \widehat{S}_{ij} \right) \left( x_{i,t+s}x_{j,t+s} - \widehat{S}_{ij} \right),$$

$$\text{Var} \left( \mathcal{S}_{ij} \right)^{\text{San},b} := \frac{1}{n} \left( \mathcal{Y}_{ij}^{\text{San}}(0) + 2 \sum_{s=1}^{n-1} \kappa(s/b) \mathcal{Y}_{ij}^{\text{San}}(s) \right), \quad b > 0,$$

\(^5\)Calculated by resampling.
where $\kappa$ is a kernel which has to fulfill Assumption B in Andrews (1991). In the following, the analysis is restricted to the truncated kernel $\kappa_{\text{TR}}(x) = \{ 1 \text{ for } |x| \leq 1, 0 \text{ otherwise} \}$ to obtain less cluttered formulas. The kernel parameter $b$ describes how many time lags are taken into account. The Sancetta-Estimator behaves well in the LDL: the main theoretical result states that for

(i) a fixed decay of the autocorrelation,

(ii) $b,n \to \infty$,

(iii) $b^2$ increasing at a lower rate than $n$,

the estimator is consistent independently of the concentration $c$ (for details, see (Sancetta, 2008)). This is in line with the results on the i.i.d. case in (Ledoit & Wolf, 2004; Chen et al., 2010; Bartz & Müller, 2013): as long as $n$ increases, all of these shrinkage estimators are consistent.

Bias of the Sancetta-Estimator In the following, it is shown that the Sancetta-Estimator is suboptimal in the case of finite sample sizes: it has a non-negligible bias. To understand this, consider a lag $s$ large enough to have $|y_{ij}(s)| \approx 0$. Approximating the expectation of the Sancetta-Estimator, it gets clear that it has a negative bias:

$$\mathbb{E}\left[ \hat{S}_{ij}^{\text{San}}(s) \right] \approx \mathbb{E}\left[ \frac{1}{n} \sum_{t=1}^{n-s} \left( x_{it}x_{jt}x_{i,t+s}x_{j,t+s} - \bar{S}_{ij}^2 \right) \right].$$

$$\approx \frac{n-s}{n} \mathbb{E}^2[\bar{S}_{ij}] - \mathbb{E}[\bar{S}_{ij}^2] = -\frac{n-s}{n} \mathbb{V} \bar{S}_{ij} < 0.$$

Bias-corrected (BC) estimator I propose a bias-corrected estimator for the variance of the entries in the sample covariance matrix:

$$\hat{\bar{S}}_{ij}^{\text{BC}}(s) := \frac{1}{n} \sum_{t=1}^{n-s} \left( x_{it}x_{jt}x_{i,t+s}x_{j,t+s} - \bar{S}_{ij}^2 \right),$$

$$\mathbb{V}[\bar{S}_{ij}]^{\text{BC},b} := \frac{1}{n-1 - 2b + b(b+1)/n} \left( \hat{\bar{S}}_{ij}^{\text{BC}}(0) + 2 \sum_{s=1}^{n-1} \kappa_{\text{TR}}(s/b) \hat{\bar{S}}_{ij}^{\text{BC}}(s) \right), \quad b > 0.$$  

The estimator $\hat{\bar{S}}_{ij}^{\text{BC}}(s)$ is very similar to $\hat{S}_{ij}^{\text{San}}(s)$, but slightly easier to compute. The main difference is the denominator in $\mathbb{V}[\bar{S}_{ij}]^{\text{BC},b}$: it is smaller than $n$ and thus corrects the negative bias.

3.3.2 Theoretical results

It is straightforward to extend the theoretical results on the Sancetta-Estimator (Sancetta (2008), see summary above) to the proposed BC-Estimator. In the following, to better understand the limitations of the Sancetta-Estimator, I provide a complementary theoretical analysis in the FOLDL.

Assumption on the uncorrelatedness of time-lagged higher moments The analysis is based on the assumptions discussed in the last section. For the analysis in the presence of autocorrelations, (A3) and (A3’) have to be extended to incorporate time-lags:

$$(A3) \quad \forall s : \sum_{i,j,kl \in Q_p} \frac{\text{Cov}[y_{ij}^p, y_{k,l+1}^p]}{|Q_p|} = O(p^{-1}),$$

$$(A3’) \quad \forall s : \sum_{i,j,kl \in Q_p} \text{Cov} \left[ y_{ij}^p y_{k,l+1}^p, (y_{k,l+1}^p y_{k,l+1+1}^p) \right] = O(p^{-1}).$$

In his simulations, Sancetta uses the Bartlett kernel. For fixed $b$, this increases the truncation bias.
Remarks on the assumptions As above, the restriction on the eighth moment (assumption (A2)) is necessary because the estimators eq. (3.12), (3.13), (3.15) and (3.16) contain fourth moments, their variances therefore contain eighth moments. Note that here I use the weaker restriction on the observations instead of the assumption (LW2) in the eigenbasis (see last section): (A2) imposes no restriction on the covariance structure.

To quantify the effect of averaging over dimensions, assumption (AC3) and (AC3') restrict the correlations of higher moments in the eigenbasis. As discussed above, this assumption is trivially fulfilled for Gaussian data, but much weaker.

Assumption (A5) rules out the degenerate case of adding observation channels without any variance while assumption (A6) and (A7) exclude distributions with arbitrarily heavy tails.

Based on these assumptions, it is possible to analyze the difference between the Sancetta-Estimator and the proposed estimator for large $p$:

**Theorem 3.11** (FOLDL consistency under AC). Let (A2), (AC3), (AC3'), (A5), (A6) and (A7) hold. We then have

\[
\frac{1}{p^2} \sum_{ij} \text{Var} \left( \tilde{S}_{ij} \right) = \Theta(1) \tag{3.17}
\]

\[
\mathbb{E} \left[ \frac{1}{p^2} \sum_{ij} \left( \text{Var}_{San,b} \left( \tilde{S}_{ij} \right) - \text{Var} \left( \tilde{S}_{ij} \right) \right) \right]^2 = \left( \text{Bias}_{San,b} + \text{Bias}_{TR}^{San,b} \right)^2 + O \left( \sum_{j} \gamma_j^2 \right) \tag{3.18}
\]

\[
\mathbb{E} \left[ \frac{1}{p^2} \sum_{ij} \left( \text{Var}_{BC,b} \left( \tilde{S}_{ij} \right) - \text{Var} \left( \tilde{S}_{ij} \right) \right) \right]^2 = \left( \text{Bias}_{TR}^{BC,b} \right)^2 + O \left( \sum_{j} \gamma_j^2 \right) \tag{3.19}
\]

where the $\gamma_i$ denote the eigenvalues of $\mathbf{C}$ and

\[
\text{Bias}_{San,b} := -\frac{1}{p^2} \sum_{ij} \left\{ \frac{1 + 2b - b(b + 1)}{n} \text{Var} \left( \tilde{S}_{ij} \right) - \frac{4}{p^3} \sum_{s=1}^{b} \sum_{t=n-s}^{n} \sum_{u=1}^{n} \text{Cov} \left[ x_{it} x_{jt}, x_{iu} x_{ju} \right] \right\} \tag{3.20}
\]

\[
\text{Bias}_{TR}^{San,b} := -\frac{2}{p^2 n} \sum_{ij} \sum_{s=b+1}^{n} \frac{n-s}{n} \text{Cov} \left[ x_{it} x_{jt}, x_{i,t+s} x_{j,t+s} \right] \tag{3.21}
\]

\[
\text{Bias}_{TR}^{BC,b} := -\frac{2}{p^2 n} \sum_{ij} \sum_{s=b+1}^{n} \frac{n-s}{n} \text{Cov} \left[ x_{it} x_{jt}, x_{i,t+s} x_{j,t+s} \right] \tag{3.22}
\]

**Proof.** see appendix C.1.

Remarks on Theorem 3.11 (i) The squared error of both estimators consists of a bias and a variance term. Both estimators have a truncation bias which is a consequence of including only a limited number of time lags into the variance estimation. When $b$ is chosen sufficiently high, this term gets close to zero.

(ii) The Sancetta-Estimator has an additional bias term which is smaller than zero in each dimension and therefore does not average out. Simulations below show that, as a consequence, the Sancetta-Estimator has a strong bias which gets larger with increasing lag parameter $b$.

(iii) The variance of both estimators behaves as $O\left( \sum_{i} \gamma_i^2 \left/ \left( \sum_{i} \gamma_i \right)^3 \right. \right)$: the more the variance of the data is spread over the eigendirections, the smaller the variance of the estimators. This bound is minimal if the eigenvalues are identical.

(iv) Theorem 3.11 does not make a statement on the relative sizes of the variances of the estimators. Note that the BC estimator mainly differs by a multiplicative factor $> 1$, hence the variance is larger, but not relative to the expectation of the estimator.
3.3. COVARIANCE SHRINKAGE FOR AUTOCORRELATED DATA

Figure 3.12. Dependence of the variance estimates on the dimensionality. Averaged over $R = 50$ models. $n = 250$.

3.3.3 Simulations

The simulations here are based on those in Sancetta (2008): results are averaged over $R = 50$ multivariate Gaussian AR(1) models

$$x_t = Ax_{t-1} + \epsilon_t,$$

with parameter matrix $A = \psi_{AC} \cdot I$, with $\psi_{no AC} = 0$, $\psi_{low AC} = 0.7$, and $\psi_{high AC} = 0.95$ (see Figure 3.10). The innovations $\epsilon_{it}$ are Gaussian with variances $\sigma_i^2$ drawn from a log-normal distribution with mean $\mu = 1$ and scale parameter $\sigma = 0.5$. For each model, $K = 50$ data sets are generated to calculate the standard deviations of the estimators and to obtain an approximation of $p^{-2} \sum_{ij} \text{Var}(S_{ij})$.

Simulation 1 In the first simulation, the dependency of the estimators on the dimensionality of the data is analyzed. The number of observations is fixed at $n = 250$ and the lag parameter $b$ chosen by hand such that the whole autocorrelation is covered: $b_{no AC} = 10$, $b_{low AC} = 20$ and $b_{high AC} = 90$. Figure 3.12 shows that the LW-Shrinkage variance estimator is unbiased and has low variance in the no AC-setting, but under the presence of autocorrelation it strongly underestimates the variance. As predicted by Theorem 3.11, the Sancetta-Estimator is also biased; its bias stays constant for increasing dimensionality. The proposed estimator has no visible bias. For increasing dimensionality the variances of all estimators decrease. Relative to the average estimate, there is no visible difference between the standard deviations of Sancetta- and the BC-Estimator.

Simulation 2 The second simulation analyzes the dependency on the lag parameter $b$ for fixed dimensionality $p = 200$ and number of observations $n = 250$. In addition to variance estimates, Figure 3.13 reports shrinkage intensities and PRIALs. The three quantities show very similar behavior. LW-Shrinkage performs well in the no AC-case, but is strongly biased in the autocorrelated settings. The Sancetta-Estimator is very sensitive to the choice of the lag parameter $b$. For low autocorrelation, the bias at the optimal $b$ is small: only a small number of biased terms is included. For high autocorrelation the optimal $b$ is larger, the higher number of biased terms causes a larger bias. The BC-Estimator is very robust: it performs well for all $b$ large enough to capture the autocorrelation. For very large $b$ its variance increases slightly, but this has practically no effect on the PRIAL.

An interesting aspect is that the BC-Estimator even outperforms shrinkage based on the population $\text{Var}(\hat{S}_{ij})$ (calculated by resampling). This results from the correlation of the estimator $\text{Var}(\hat{S}_{ij})_{BC,b}$ with the sample estimate eq. (3.13) of the denominator in eq. (3.5).

---

7 More complex parameter matrices or a different generative model do not pose a problem for the bias-corrected estimator. The simple model was chosen for clarity of presentation.

8 For $b < 1$, optimal in the no AC-setting, the Sancetta and the BC-estimator are equivalent to standard LW-Shrinkage.
Figure 3.13. Robustness to the choice of lag parameter \( b \). Variance estimates (upper row), shrinkage intensities (middle row) and improvement over sample covariance (lower row). Averaged over \( R = 50 \) models. \( p = 200, n = 250 \).

### 3.3.4 Real world data: BCI based on motor imagery

As an example for highly autocorrelated data serves the MI-BCI data set described in Section 1.6. MI-BCI is a typical application scenario for Common Spatial Patterns (CSP) (Lotte et al., 2007; Blankertz et al., 2008): CSP is used for dimension reduction in classification settings where (i) each datapoint is a time series of observations and (ii) the discriminative information between two classes lies in the signal variance. Then CSP yields filters for the classes A and B which are defined by the directions where the ratio of the variances is maximal:

\[
\hat{f}^{A/B}_i := \arg \max_{f \in \mathbb{F}, \mathbb{f}^{A/B}_i < i} \frac{f^T \hat{C}^{A/B}_i f}{f^T (\hat{C}^A + \hat{C}^B) f}.
\]

As common in Brain-Computer Interfacing, an LDA classifier is trained on features \( x^{\text{CSP}}_i = \log \left( \hat{\text{Var}}(X \hat{f}_i) \right) \).

For the analysis, the frequency band was optimized for each subject and from the class-wise covariance matrices, 1-3 filters per class were extracted by Common Spatial Patterns (CSP), adaptively chosen by a heuristic (see Blankertz et al. (2008)). Linear Discriminant Analysis is trained on log-variance features.

To improve the estimation of the class covariances on this highly autocorrelated data set, LW-Shrinkage, Sancetta-Shrinkage, cross-validation and and the proposed BC-Shrinkage were applied. The
3.3. COVARIANCE SHRINKAGE FOR AUTOCORRELATED DATA

Figure 3.14. BCI motor imagery data for lag parameter \( b = 75 \) (upper row) and \( b = 300 \) (lower row). Averaged over all 80 subjects and \( K = 100 \) runs.

As discussed in this chapter, LW-Shrinkage is dominated by the influence of high-variance directions (Bartz & Müller, 2013), which are pronounced in this data set. To reduce this effect, only for the calculation of the shrinkage intensities, the first five principal components are rescaled to have the same variance as the sixth principal component. The application of aoc-Shrinkage is not feasible, as aoc-Shrinkage is not invariant to autocorrelations.

The dependency of the four algorithms on the number of supplied training trials is analyzed. Figure 3.14 (upper row) shows results for an optimized time lag \( (b = 75) \) which captures the autocorrelation of the data well (outer left). Taking the autocorrelation into account makes a clear difference (middle left/right): while LW-Shrinkage outperforms the sample covariance, it is clearly outperformed by the autocorrelation-adjusted approaches. The Sancetta-Estimator is slightly worse than the proposed estimator. The shrinkage intensities (outer right) are extremely low for LW-Shrinkage and the negative bias of the Sancetta-Estimator clearly shows up for small numbers of training trials.

Figure 3.14 (lower row) shows results for a too large time lag \( (b = 300) \). The performance of Sancetta-
Shrinkage strongly degrades as its shrinkage intensities get smaller, while the proposed BC-Shrinkage is robust to the choice of $b$. Only for the smallest number of trials a small degradation of performance is observed. Figure 3.15 compares the BC-Estimator to the four other approaches for 10 training trials: it significantly outperforms LW- and Sancetta-Shrinkage for both the larger ($b = 300$, $p \leq 0.01$) and the smaller time lag ($b = 75$, $p \leq 0.05$).

All analytic shrinkage procedures minimize the expected squared error of the covariance matrix, while cross-validation directly optimizes the classification performance (see Section 3.1). Yet, Figure 3.14 (middle) shows that for small numbers of training trials the proposed estimator outperforms CV-Shrinkage, although the difference is not significant (see Figure 3.15). For larger numbers of training trials CV-Shrinkage performs better. This shows that the ESE is not a very good proxy for classification accuracies in the context of CSP: for optimal ESE, shrinkage intensities decrease with increasing number of observations. Shrinkage intensities in CV-Shrinkage instead stay on a constant level between 0.1 and 0.15. Figure 3.15 (right) shows that the three analytic shrinkage approaches ($b = 300$) have a huge advantage over cross-validation (10 folds/10 parameter candidates) with respect to runtime.

3.4 Chapter summary

Regularized covariance estimators are a good and popular alternative to restricted estimators. In particular, the LW-Shrinkage estimator is widely used because it yields good performance, has no free parameters besides the choice of target and is very fast compared to cross-validation.

Shrinkage and structure In this chapter, I have clarified the limits of the applicability of the analytic shrinkage formula by Ledoit and Wolf. I showed that its assumptions are often violated in practice since real world data sets tend to exhibit highly dependent observations.

I extended shrinkage theory to these cases, providing proofs of consistency and showing that for typical covariance structures in real world data the shrinkage intensity converges to zero in the limit of infinitely many observations and dimensions. In addition, I showed that under certain conditions consistency even holds when only the number of dimensions goes to infinity.

Consequently, I have proposed an algorithm which automatically restricts shrinkage to the orthogonal complement of the strongest eigendirections if this leads to an improvement in mean squared error. Simulations show that (i) restricting shrinkage to the orthogonal complement can reduce the error in covariance estimation and (ii) the proposed automatic method correctly identifies the number of eigendirections which should be excluded from shrinkage.

On real world data from the domains of finance, spoken letter and optical character recognition, and neuroscience, the proposed method yields improved robustness to noise and a significant performance enhancement.

The aoc-Shrinkage approach is related to factor modeling. Restricting shrinkage to the orthogonal complement can be seen as a kind of hybrid between shrinkage and factor modeling. A drawback of aoc-Shrinkage is that its implementation is slightly more involved and that its calculation requires an eigendecomposition.

Shrinkage and autocorrelation While Ledoit-Wolf Shrinkage assumes i.i.d. data, many real world data sets, for example from video, audio, finance, biomedical engineering or energy systems clearly violate this assumption as strong autocorrelation is present. Intuitively this means that the information content per data point is reduced, and thus the covariance estimation problem becomes harder: the dimensionality remains unchanged but the effective number of available samples decreases. Thus stronger regularization is required and LW-Shrinkage needs to be adjusted.

Sancetta (2008) already took the first step into this important direction by providing a LDL-consistent estimator under i.i.d. violations. In this chapter I analyzed finite sample sizes and showed that (i) even apart from truncation bias—which results from including a limited number of time lags—Sancetta's
estimator is biased, (ii) this bias is only negligible if the autocorrelation decays fast compared to the length of the time series and (iii) the Sancetta-Estimator is very sensitive to the choice of lag parameter.

I proposed an alternative estimator which is (i) both consistent and —apart from truncation bias— unbiased and (ii) highly robust to the choice of lag parameter: in simulations on toy and real world data, the proposed estimator yielded large improvements for small samples and/or suboptimal lag parameter. Even for optimal lag parameter there is a slight but significant improvement.

Analyzing data from a motor imagery BCI experiment, it can be seen that (i) the BCI data set possesses significant autocorrelation, that (ii) this adversely affects CSP based on the sample covariance and LW-Shrinkage (iii) this effect can be alleviated using the novel estimator, which is shown to (iv) compare favorably to Sancetta’s estimator.
Chapter 4

Regularized Estimation II: MTS

This chapter detours from pure covariance matrix estimation. The shrinkage concept is not only applicable to covariance matrices: apart from sample covariances, shrinkage has been successfully applied to sample means (Stein, 1956), wavelets (Donoho & Johnstone, 1995), density estimators (Sancetta, 2013) and kernels (Muandet et al., 2014).

I extend the shrinkage framework to multiple targets by allowing simultaneous shrinkage to a set of targets. Application scenarios include settings with (i) additional data sets from potentially similar distributions, (ii) non-stationarity, (iii) a natural grouping of the data or (iv) multiple alternative estimators which could serve as targets.

I show that this Multi-Target Shrinkage (MTS) can be translated into a quadratic program and derive conditions under which the estimation of the shrinkage intensities yields optimal expected squared error in the limit. For the sample mean and the sample covariance as specific instances, and for the LDL as well as for the FOLDL, I derive conditions under which the optimality of MTS is applicable. I then show the effectiveness for the estimation of both the mean and the covariance in extensive simulations and on real world data. This chapter is based on an article submitted to the Journal of Machine Learning Research (Bartz et al., 2014).

MTS was developed in parallel by Lancewicki & Aladjem (2014), with a clear focus on covariances and the construction of targets. In contrast, this chapter presents a more complete picture: MTS is introduced in a general, estimator-independent framework. In addition, a theoretical analysis of consistency is provided, the relationship to transfer learning is discussed and the first application to real world data is shown.

Chapter outline Section 4.1 gives an intuitive introduction to MTS. After defining some notation in Section 4.2, in Section 4.3 a quadratic program for the vector of optimal shrinkage intensities is derived and its convergence is proven. Section 4.4 and Section 4.5 show that the general theory can be applied to the sample mean and the sample covariance. The effectiveness in simulations and on real world data is shown in Sections 4.6 and 4.7, respectively.

4.1 Introduction to Multi-Target Shrinkage

In the following, a generalization of the analytic shrinkage approach by Ledoit-Wolf—in the context of this chapter called Single-Target Shrinkage (STS)—to multiple shrinkage targets is proposed. Figure 4.1 illustrates Single- and Multi-Target Shrinkage (MTS) of an unbiased estimator \( \hat{\theta} \) of a parameter \( \theta \) for the

\[\text{Note that the same symbols are used for the estimator (a random variable) and the estimate (a realization of the random variable). It is clear from the context to which is referred.}\]
Figure 4.1. Geometric illustration of Multi-Target Shrinkage. The unbiased estimate and the two targets span a convex set. The optimal MTS estimate is the estimate in the convex set with minimum squared distance to the population.

Illustration on hand-written digits  As an illustration, MTS is applied to the estimation of subject-specific mean images on the data set of subject-specific handwritten digits (see Section 1.6.4). Assume that we want to estimate the mean image of digit 9 of person A from a small number of observations. In this case MTS improves over the sample mean image and STS by shrinking towards the mean images of two other subjects T1 and T2. This is seen in Figure 4.2: for MTS, the differences to the truth\(^3\) are less pronounced than in STS and the squared error is smaller.

The illustrations Figure 4.1 and 4.2 are limited to the case of simultaneous shrinkage to two shrinkage targets. MTS can handle an arbitrary number of shrinkage targets \(\hat{T}^1, \hat{T}^2, \ldots, \hat{T}^K\). Figure 4.3 shows this for the handwritten digits: as more and more targets are incorporated, the squared error decreases.

Application scenarios  There are many application scenarios for Multi-Target Shrinkage:

\(^2\)The optimum can lie on the border of the triangle if one of the targets is completely useless. Otherwise, it lies within the triangle.

\(^3\)The mean of the hold-out data for subject A serves as a proxy for the truth.
4.1. INTRODUCTION TO MULTI-TARGET SHRINKAGE

Figure 4.2. Geometric illustration of Multi-Target Shrinkage for handwritten digits. The targets are the mean images of digit 9 for two different subjects.

Figure 4.3. Decay of the squared error for increasing number of shrinkage targets. Average over $R = 10000$ random choices of digits and subjects.
• **similar data sets:** assume that $K$ additional data sets from similar distributions are available. Then, it is possible to calculate a target $\mathbf{T}^k$ on each additional data set and use MTS to decide how useful the other data sets are for the estimation task. This is a special case of transfer learning (see (Pan & Yang, 2010) for a recent review). The handwritten digits example from Figure 4.2 and 4.3 falls into this category.

• **data with group structure:** if there is a natural group structure in a data set, one can estimate $\theta$ either (a) on the whole data set or (b) for each group individually.
  
  - When $\theta$ is independent of group membership, (a) is optimal and MTS yields approximately equal weights.
  - When $\theta$ is very different for each group, (b) is optimal and MTS puts approximately no weight on the targets.
  - When $\theta$ is dependent of group membership, but similar for at least some groups, MTS provides an optimal weighting of each group which is superior to both (a) and (b).

• **non-stationarity:** assume that the parameter $\theta$ is non-stationary. MTS can yield a superior estimate of the current value of $\theta$ by treating older segments of the data as shrinkage targets.

• **multiple available biased estimators:** for covariance shrinkage, a set of biased estimators has been proposed as shrinkage targets (see Section 3.1). Which one of these structured estimators constitutes the best target depends on the structure of the population covariance matrix. The choice is based on expert knowledge or cross-validation. In contrast, MTS does not make a choice but yields an optimal weighting of all targets which is equal or superior to the optimal choice.

In the previous chapter, it was stated that the optimal STS intensity can be estimated by minimizing the ESE or by making use of a slower cross-validation approach. For MTS, the computational cost to cross-validate $K$ parameters grows with the power of $K$ which is not feasible. This motivates the extension of the approach of minimizing the ESE to multiple shrinkage targets.

### 4.2 MTS notation and distributional assumptions

Extending shrinkage to multiple targets in an estimator-independent formulation requires some additional notation and modified assumptions.

**Notation for general MTS** In Section 4.3 the general case is analyzed:

- The estimation of a set of parameters $\theta = (\theta_1, \theta_2, \ldots, \theta_q) \in \mathbb{R}^q$ is considered, for which the existence of an unbiased estimator $\tilde{\theta}$ is assumed.

- As in the last chapter, optimality is always defined w.r.t. the expected squared error (ESE), denoted by $\Delta$. The ESE of the unbiased estimator $\tilde{\theta}$, for example, is denoted by

$$\Delta^{\tilde{\theta}} := \mathbb{E}\|\tilde{\theta} - \theta\|^2.$$

**Notation for MTS of the mean and the covariance** In Sections 4.4 and 4.5, we consider the estimation of the mean and the covariance matrix, respectively. There,

- the sequence index $p$ also denotes the dimensionality of $n_p$ i.i.d. observations with mean $\mu_p$ and covariance $C_p$, given by the $(p \times n_p)$-matrix $X_p$.

- We consider $K$ additional data sets with mean $\mu^k_p$ and covariance $C^k_p$, their $n^k_p$ i.i.d. observations are given by the $(p \times n^k_p)$-matrices $X^k_p$. 

4.3 General Multi-Target Shrinkage

Table 4.1. General, mean and covariance MTS. Note that the covariance has \( p(p+1)/2 \) degrees of freedom. As for the ESE, we here consider each entry in the covariance matrix as a parameter.

<table>
<thead>
<tr>
<th>Setting</th>
<th>set of parameters</th>
<th>unbiased estimator</th>
<th>#parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>general</td>
<td>( \boldsymbol{\theta} )</td>
<td>( \hat{\theta} )</td>
<td>( q )</td>
</tr>
<tr>
<td>mean</td>
<td>( \boldsymbol{\mu} := \mathbb{E}[\mathbf{x}] )</td>
<td>( \hat{\mu} := n^{-1} \sum_i \mathbf{x}_i )</td>
<td>( q = p )</td>
</tr>
<tr>
<td>covariance</td>
<td>( \mathbf{C} := \mathbb{E}[(\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^\top] )</td>
<td>( \hat{\mathbf{C}} = \hat{\mathbf{S}} := n^{-1} \sum_i (\mathbf{x}_i - \hat{\boldsymbol{\mu}})(\mathbf{x}_i - \hat{\boldsymbol{\mu}})^\top )</td>
<td>( q = p^2 )</td>
</tr>
</tbody>
</table>

- \( \gamma_{p,1}^{(k)}, \gamma_{p,2}^{(k)}, \ldots, \gamma_{p,p}^{(k)} \) denote the eigenvalues of \( \mathbf{C}_p^{(k)} \).
- \( \mathbf{Y}_p^{(k)} = \mathbf{R}_p^{(k)\top} \mathbf{X}_p^{(k)} \) denote the observations in their respective eigenbasis, where the covariance matrices \( \mathbf{\Sigma}_p^{(k)} = \mathbf{R}_p^{(k)\top} \mathbf{C}_p \mathbf{R}_p^{(k)} \) are diagonal. The mean in the eigenbasis is denoted by \( \mathbf{\mu}_p^{Y(k)} \).
- For two data sets \( \mathbf{X}_p^{(k)} \) and \( \mathbf{X}_p^{(l)} \), we denote \( \mathbf{Z}_p^{(k)} = \mathbf{R}_p^{(l)\top} \mathbf{X}_p^{(k)} \). From the context, it will be clear which \( l \) was used to obtain \( \mathbf{Z}_p^{(k)} \).
- In the following, the sequence index \( p \) is always omitted to obtain a less cluttered notation.

Table 4.1 gives an overview of the different MTS scenarios considered in this paper.

**Distributional assumptions** The following assumptions are required:

(D1) \[ (\forall k) : \frac{1}{p} \sum_{i=1}^p \gamma_i^{(k)} = \Theta(1). \]

(D2) \[ (\forall k) \exists \gamma^{(k)} : \frac{1}{p} \sum_{i=1}^p \gamma_i^{(k)} = \Theta(p^{\gamma^{(k)}}). \]

(D3) \[ \exists \alpha_4, \beta_4 : (1 + \beta_4)\mathbb{E}^2[y_i^4] \leq \mathbb{E}[y_i^4] \leq (1 + \alpha_4)\mathbb{E}^2[y_i^4] \]

(D4) \[ \exists \alpha_8, \beta_8 : (1 + \beta_8)\mathbb{E}^2[y_i^8] \leq \mathbb{E}[y_i^8] \leq (1 + \alpha_8)\mathbb{E}^2[y_i^8] \]

The assumption (D1) states, for each data set, that for an increasing number of dimensions the variance per dimension is bounded from above and below.

The assumption (D2) restricts the dispersion of the eigenvalues: for increasing dimensionality, the dispersion is assumed to have a well-defined limit behavior. Note that (D1) implies \( 0 \leq \gamma_i^{(k)} \leq 1 \).

The assumptions (D3) and (D4) have two purposes: first they guarantee the existence of fourth and eighth moments, respectively. Second, they impose an (arbitrary) upper bound on the heaviness of the tails in the sequence \( p \).

### 4.3 General Multi-Target Shrinkage

In Single-Target Shrinkage, the linear combination of an unbiased estimator \( \hat{\theta} \) with another estimator \( \hat{T} \) (called the shrinkage target) is optimized. In the following, as in the last chapter, the linear combination is restricted to be convex:

\[
\hat{\theta}^{\text{STS}}(\lambda) := (1 - \lambda)\hat{\theta} + \lambda\hat{T}.
\]
In this chapter of the thesis, I generalize to the optimization of the convex combination of the sample covariance with a set of \( K \) targets,
\[
\tilde{\theta}^{\text{MTS}}(\lambda) := \left(1 - \sum_{k=1}^{K} \lambda_k\right) \tilde{\theta} + \sum_{k=1}^{K} \lambda_k \tilde{T}^k,
\]
where \( \lambda = (\lambda_1, \lambda_2, \ldots, \lambda_K) \in \mathbb{R}_+^K \) is subject to \( \sum_k \lambda_k \leq 1 \). The MTS objective is given by
\[
\Delta^{\text{MTS}}(\lambda) := \mathbb{E} \| \theta - \tilde{\theta}^{\text{MTS}}(\lambda) \|^2.
\]
From the MTS objective I derive a quadratic program for the optimal value of \( \lambda \):

**Theorem 4.1** (MTS quadratic program). Let the MTS quadratic program be defined by
\[
\Delta_{\text{qp}}^{\text{MTS}}(\lambda) := \frac{1}{2} \lambda^T A \lambda - b^T \lambda.
\]
with
\[
A_{kl} := \sum_{i=1}^{q} \mathbb{E} \left[ (\tilde{T}_{i}^k - \hat{\theta}_i) (\tilde{T}_{i}^l - \hat{\theta}_i) \right], \quad b_k := \sum_{i=1}^{q} \left( \text{Var}(\hat{\theta}_i) - \text{Cov}(\tilde{T}_{i}^k, \hat{\theta}_i) \right),
\]
Then it is equivalent to optimize \( \Delta^{\text{MTS}}(\lambda) \) and \( \Delta_{\text{qp}}^{\text{MTS}}(\lambda) \):
\[
\lambda^* := \arg \min_{\lambda \in \mathbb{R}_+^K, \sum_k \lambda_k \leq 1} \Delta^{\text{MTS}}(\lambda) = \arg \min_{\lambda \in \mathbb{R}_+^K, \sum_k \lambda_k \leq 1} \Delta_{\text{qp}}^{\text{MTS}}(\lambda).
\]

**Proof.** see appendix D.1.

The quadratic program is governed by the parameters \( A \) and \( b \), quantifying the quality of the targets and the unbiased estimator, respectively. The vector \( b \) contains the variance of the unbiased estimator, adjusted for correlation with the targets. The diagonal elements in the matrix \( A \) contain information on the variance and bias of the targets and on the correlation with the unbiased estimator. A target \( \tilde{T}^k \) is useful if the entry in \( A_{kk} \) is small relative to the variance of the unbiased estimator. The off-diagonal elements in the matrix \( A \) contain information on the correlation between targets.

### 4.3.1 Estimation of Multi-Target Shrinkage

The optimal shrinkage intensities \( \lambda^* \) depend on the unknown parameters \( A \) and \( b \) of the quadratic program eq. (4.4). I propose the following estimators:
\[
\hat{\lambda} := \arg \min_{\lambda \in \mathbb{R}_+^K, \sum_k \lambda_k \leq 1} \hat{\Delta}^{\text{MTS}}(\lambda), \quad \hat{\Delta}^{\text{MTS}}(\lambda) := \frac{1}{2} \lambda^T \hat{A} \lambda - \hat{b}^T \lambda \quad \text{with}
\]
\[
\hat{A}_{kl} := \sum_{i=1}^{q} \left( \tilde{T}_{i}^k - \hat{\theta}_i \right) \left( \tilde{T}_{i}^l - \hat{\theta}_i \right), \quad \hat{b}_k := \sum_{i=1}^{q} \left\{ \text{Var}(\hat{\theta}_i) - \text{Cov}(\tilde{T}_{i}^k, \hat{\theta}_i) \right\},
\]
where the unbiased estimator \( \hat{\theta} \), the targets \( \tilde{T}^k \) and the estimators of variance and covariance appearing in \( \hat{b} \) depend on the application scenario.

For a general parameter set \( \theta \), the following theorem relates the limit behavior of the estimators in \( \hat{b} \) and of linear combinations of the estimators in \( \hat{A} \) to the limit behavior of \( \Delta^{\text{MTS}}(\lambda) \) and \( \lambda^* \):

---

4 Setting \( \tilde{T}^{K+1} = 0 \) and allowing for \( \lambda \in \mathbb{R}^{K+1} \), this turns into an arbitrary linear combination which can deal with arbitrarily rescaled targets. Theoretical results can be extended at the cost of clarity and accessibility.
4.3. GENERAL MULTI-TARGET SHRINKAGE

**Theorem 4.2. (consistency of MTS)** Let us assume a sequence of models indexed by $p$ such that

\begin{align*}
\text{(G1)} & \quad \exists \tau : \Delta = \Theta (p^{r_\lambda}), \\
\text{(G2)} & \quad \forall k \exists \tau_A : A_{kk} = \Theta (p^{r_\lambda}), \quad \forall k : b_k = \Theta (p^{r_\lambda}) \\
\text{(G3)} & \quad \| \hat{A}_{kl} - A_{kl} \| = o \left( p^{0.5(r_\lambda + r_\lambda^k)} \right), \quad \| \hat{b}_k - b_k \| = o (p^{r_\lambda}) \\
\text{(G4)} & \quad \forall k : \min_{\alpha \in \mathbb{R}^K} \sum_{i=1}^q \mathbb{E} \left[ \left( \sum_{i=1}^K \alpha_i (\hat{T}_i^k - \hat{\theta}_i) \right)^2 \right] = \Theta \left( p^{r_\lambda} \right)
\end{align*}

We then have

(i) \quad \forall k : \lambda^*_k, \hat{\lambda}_k = O \left( p^{(r_\theta - r_\lambda^k)/2} \right).

(ii) \quad \frac{\Delta^{\text{MTS}}(\hat{\lambda}) - \Delta^{\text{MTS}}(\lambda^*)}{\Delta^\theta} = o(1)

If one strengthens (G4) to hold $\forall \alpha \in \mathbb{R}^K$, we also have

(iii) \quad \| \lambda^* - \hat{\lambda} \| = o(1)

**Proof.** see appendix D.2.

The assumptions (G1) and (G2) state that all estimators have a well-defined limit behavior w.r.t. the ESE. Note that assuming that $\Delta^\theta$ and $b_k^\theta$ have the same limit behavior implies that none of the targets is identical to the unbiased estimator.

Assumption (G3) states that the relative errors in the entries of the estimators $\hat{A}_{kl}$ and $\hat{b}_k$ go to zero in the limit. In the following, this property is called **consistency of $\hat{A}$ and $\hat{b}$**.

Assumption (G4) states that the linear combination of a set of targets cannot have better limit behavior w.r.t. the ESE than the best single target in the set. This assumption is necessary because linear dependence of targets can result in $A$ having small eigenvalues for which the relative error does not go to zero.

To illustrate the assumptions consider the handwritten digits example. A hypothetical sequence of models would consist of images with increasing resolution ($p \times p$ pixels) and an increasing number of observations for each subject. In such a sequence, the ESE of the sample estimator for subject $A$ would have a clear limit behavior and hence fulfill (G1). The similarity between the digits of subjects $A$ and e.g. $T_1$ defines the similarity of the images. Hence a clear limit behavior of $A$ (G2) is to be expected. With increasing $p$ and $n$, a better estimation of the variance of the sample mean and the similarity between subjects is possible, and hence the relative errors in $b$ and $A$ would go to zero (G3). Two subjects $T_1$ and $T_2$ whose differences to subject $A$ exactly cancel out in a linear combination would violate Assumption (G4). This is highly unlikely.

Part (i) of Theorem 4.2 states that a target $T^k$ which has worse limit behavior w.r.t. the ESE than the sample estimator $\hat{\theta}$ does not contribute in the limit.

Part (ii) is the most important result. It states that the ESE of the MTS estimator $\hat{\lambda}$ (normalized by the error of the sample estimator) converges to the ESE of the optimal $\lambda^*$\(^6\). In the following, this property is called **consistency of MTS**.

Part (iii) shows that $\lambda^*$ is, under a restriction on the linear dependency of the targets, identifiable and that the estimator $\hat{\lambda}$ converges to $\lambda^*$. In the following, this property is called **consistency of the estimator $\hat{\lambda}$**.

---

\(^5\)For an off-diagonal element $A_{kl}$, we consider the error relative to $\sqrt{A_{kk}A_{ll}}$.

\(^6\)Note that $\frac{\Delta^{\text{MTS}}(\hat{\lambda}) - \Delta^{\text{MTS}}(\lambda^*)}{\Delta^{\text{MTS}}(\lambda^*)} = o(1)$ does not hold in general.
4.4 Multi-Target Shrinkage of the mean

In this section, the MTS approach is applied to the \( p \)-dimensional sample mean:

\[
\theta = \mu, \quad \hat{\theta} = \hat{\mu} = (\hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_{q=p}).
\]

As shrinkage targets, a set of sample means \( \hat{\mu}_1, \hat{\mu}_2, \ldots, \hat{\mu}_K \) of additional data sets \( X^1, X^2, \ldots, X^K \) is taken, drawn from potentially different distributions. This yields

\[
A_{kl} = \sum_{i=1}^{p} \mathbb{E} \left[ (\hat{\mu}_i^k - \hat{\mu}_i) (\hat{\mu}_i^l - \hat{\mu}_i) \right] \quad b_k = \sum_{i=1}^{p} \{ \text{Var}(\hat{\mu}_i) - \text{Cov}(\hat{\mu}_i^k, \hat{\mu}_i) \}. \quad (4.7)
\]

Cov(\( \hat{\mu}_i^k, \hat{\mu}_i \)) = 0 holds and sample estimates for \( \hat{A} \) and \( \hat{b} \) are given by

\[
\hat{A}_{kl} := \sum_{i=1}^{p} (\hat{\mu}_i^k - \hat{\mu}_i) (\hat{\mu}_i^l - \hat{\mu}_i) \quad \hat{b}_k := \hat{b} := \sum_{i=1}^{p} \hat{\text{Var}}(\hat{\mu}_i), \quad (4.8)
\]

where the estimator of the variance of the sample mean is defined by

\[
\hat{\text{Var}}(\hat{\mu}_i) := \frac{1}{n(n-1)} \sum_{t=1}^{n} (x_{it} - \hat{\mu}_i)^2.
\]

Remark MTS of the mean can be interpreted as a weighting of each data point. Data points in \( X \) are weighted by \( (1 - \sum_{l=1}^{K} \lambda_l^k)n^{-1} \) and data points in \( X^k \) are weighted by \( \lambda_k^k n_k^{-1} \).

Assuming that the distributions of the data sets only differ with respect to their means, the optimal weight of each original data point is larger than or equal to the weight of the data points from the additional data sets. This translates into a constraint on the quadratic program:

\[
\forall k : \quad \lambda_k^k n_k^{-1} \leq \left( 1 - \sum_{i=1}^{K} \lambda_i^k \right) n^{-1}.
\]

In many applications it is reasonable to impose this constraint to increase the precision, in particular for small sample sizes.

4.4.1 Consistency of MTS of the mean

In this section, the conditions under which MTS of the mean is consistent are established by deriving necessary conditions for the estimators eq. (4.8) to fulfill the assumptions of Theorem 4.2. This is shown for both asymptotic settings.

LDL consistency of MTS of the mean Let us first consider the LDL.

Theorem 4.3 (LDL consistency of MTS of the mean). Let us assume a sequence of statistical models indexed by \( p \) for which (D1), (D2), (D3) and

\[
(M1) \quad \forall k \exists \tau_k^\mu \leq 1 : ||\mu^k - \mu||^2 = \Theta(p^\tau_k^\mu),
\]

\[
(M2) \quad \forall k : \quad \tau_k^\mu < 2 \max(0, \tau_k^\mu) + 1 \quad \text{and} \quad \tau_k^\mu < 2 \max(0, \min_k \tau_k^\mu) + 1
\]

\[
(M3) \quad \forall k |\tau_k^\mu > 1 : \min_{\alpha \in \mathbb{R}_{p>0}} \left\{ \sum_{i=1}^{n} \alpha_i (\mu^i - \mu) \right\}^2 = \Theta \left( p^{\tau_k^\mu} \right)
\]
4.5. MULTI-TARGET SHRINKAGE OF THE COVARIANCE MATRIX

hold.

Then assumptions (G1), (G2), (G3), and (G4) of Theorem 4.2 are fulfilled, MTS of the mean is consistent and

\[ \forall k : \tau_k^k > 0 : \lambda_k^* = \lambda_k = \mathcal{O} \left( p^{-\tau_k^k/2} \right) \]

holds. If (M3) holds for \( \alpha \in \mathbb{R}^K \), \( \lambda^* \) is identifiable and \( \lambda \) is consistent.

Proof. see appendix D.3.

Assumption (M1) states that the distance between data and target mean needs to have a clear limit behavior. Unrealistic sequences of models with \( \tau_k^k > 1 \), in which the distance between data and target mean grows faster than the dimensionality, are excluded.

Assumption (M2) limits the eigenvalue dispersion of the data sets in dependence on the distance between data and target mean. Intuitively, if there are strong directions whose contributions are at a constant level independent of \( p \) and hence do not average out, small distances between data and target mean cannot be estimated reliably.

Assumption (M3) states that for no linear combination of target means the error decreases at a faster rate than for all included single target means.

Theorem 4.3 states conditions under which MTS of the mean is consistent in the LDL. In addition it states that data sets with increasing mean distance (\( \tau_k^k > 0 \) in (M1)) do not contribute to the MTS estimate in the LDL limit: for \( n \to \infty \), the sample means \( \hat{\mu}_k \) of these data sets do not constitute useful shrinkage targets because the limit behaviour of the sample mean \( \hat{\mu} \) is superior.

FOLDL consistency of MTS of the mean Let us now consider the case where only the dimensionality \( p \) goes to infinity, while \( n \) remains constant.

Theorem 4.4 (FOLDL consistency of MTS of the mean). Let us assume a sequence of statistical models indexed by \( p \) for which (D1), (D2), (D3), assumption (M1) from Theorem 4.3 and

\[ (\forall k) \quad \tau_k^k < 1 \quad \text{and} \quad \tau_\gamma < 1 \]

\[ (\forall k) \quad \sum_{i,j \neq i} \text{Cov} \left( \frac{y_i^{(k)}}{y_i^{(k)}} \right) = o \left( p^2 \right) \]

hold. Then assumptions (G1), (G2), (G3), and (G4) of Theorem 4.2 are fulfilled and MTS of the mean is consistent and \( \lambda \) is a consistent estimator.

Proof. see appendix D.4.

In the FOLDL, consistency results from averaging over dimensions. Therefore, consistency requires stronger restrictions on the correlation between dimensions. Assumption (M2') states that the dispersion of the eigenvalues (D2) has to grow slower than \( \Theta(p) \). Otherwise, strong eigendirections exist whose influence on the MTS estimate remains at a constant level in the sequence of models. Assumption (M4) states that the correlation between squared uncorrelated variables, on average, converges to zero.

Note that identifiability holds even without Assumption (M3).

4.5 Multi-Target Shrinkage of the covariance matrix

The second application of MTS considers the sample covariance matrix as an estimator of the population covariance matrix:

\[ \theta = C, \quad \hat{\theta} = \hat{S}, \quad \hat{S}_{ij} = n^{-1} \sum_{s=1}^{n} (x_{is} - \hat{\mu}_i)(x_{js} - \hat{\mu}_j). \]
For the sample covariance matrix, two classes of targets are considered:

- as for the sample mean, it is possible to shrink to a set of sample covariance matrices $\hat{S}^1, \ldots, \hat{S}^K$ from additional data sets $X^1, X^2, \ldots, X^K$.
- several biased estimators $\hat{C}^1, \hat{C}^2, \ldots, \hat{C}^K$ of the sample covariance matrix exist which can be used as targets (see Section 3.1).

In total, one obtains a set of targets $\hat{T}^1, \hat{T}^2, \ldots, \hat{T}^K$ with

$$
A_{kl} = \sum_{i,j=1}^{p} E \left[ (\hat{T}_{ij} - \hat{S}_{ij}) (\hat{T}_{lj} - \hat{S}_{lj}) \right] \quad \text{and} \quad b_k = \sum_{i,j=1}^{p} \{ \text{Var}(\hat{S}_{ij}) - \text{Cov}(\hat{T}_{ij}, \hat{S}_{ij}) \}. 
$$

Sample estimates $\hat{A}$ and $\hat{b}$ are given by

$$
\hat{A}_{kl} = \sum_{i,j=1}^{p} (\hat{T}_{ij} - \hat{S}_{ij}) (\hat{T}_{lj} - \hat{S}_{lj}) \quad \text{and} \quad \hat{b}_k = \sum_{i,j=1}^{p} \text{Var}(\hat{S}_{ij}),
$$

where the estimator of the variance of the sample covariance is given by

$$
\hat{\text{Var}}(\hat{S}_{ii}) := \frac{1}{(n-1)n} \sum_{s} \left( x_{is}x_{js} - \frac{1}{n} \sum_{t} x_{it}x_{jt} \right)^2. 
$$

To keep the notation simple, $\forall k : \mu = \mu^k = 0$ is assumed.

### 4.5.1 Consistency of MTS of the covariance

In this section, the conditions under which MTS of the covariance is consistent are established by showing what is required for the estimators eq. (4.9) to fulfill the assumptions of Theorem 4.2. As in the previous section, both asymptotic settings are considered.

**Theorem 4.5** (LDL consistency of MTS of the covariance). Let us assume a sequence of statistical models indexed by $p$ for which (D1), (D2), (D3), (D4) and

\begin{align*}
(\text{C1}) & \quad \forall k \exists \tau^k_C \leq 2 : ||C^k - C||^2 = \Theta(p^{\tau^k_C}), \\
(\text{C2}) & \quad \sum_{i,j,k,l \in Q_p} \frac{\text{Cov} [y_{i1}y_{j1}, y_{k1}y_{l1}]}{|Q_p|} = o(1)
\end{align*}

where $Q_p$ is the set of all quadruples consisting of distinct integers between 1 and $p$,

\begin{align*}
(\text{C3}) & \quad 1 + 2 \tau^{(k)}_{\gamma} < 2 \max(1, \min_{k} \tau^k_C), \\
(\text{C4}) & \quad \forall k |r^k_C > 1 : \min_{\alpha \in R^K} \left\| \sum_{t} \alpha_t (C^t - C) \right\|^2 = \Theta \left( p^{r^k_C} \right)
\end{align*}

hold. Then, for the set of targets in (Schrëfer & Strimmer, 2005) and targets given by additional data sets, assumptions (G1), (G2), (G3) and (G4) of Theorem 4.2 are fulfilled. Hence MTS of the covariance is consistent and

$$
\forall k |r^k_C > 1 : \lambda^*_k, \hat{\lambda}_k = O \left( p^{(1-r^k_C)/2} \right)
$$

holds. If (C4) holds for $\alpha \in R^K$, $\lambda^*$ is identifiable and $\hat{\lambda}$ is consistent.
Proof. see appendix D.5.

Assumption (C1) states that the distance of the data covariance matrices to each target covariance needs to have a clear limit behavior. It excludes unrealistic sequences of models with $\tau^C_k > 2$ in which the distance between data and target grows faster than the number of entries in $C$.

Assumption (C2) restricts the average covariance between products of uncorrelated variables. This assumption is quite weak (compare to (Ledoit & Wolf, 2004)).

Assumption (C3) limits the eigenvalue dispersion of the data sets in dependence of the distance between data and target covariance. This is analogous to assumption (M2) for MTS of the mean.

Assumption (C4) states that for no linear combination of target covariances the error decreases at a faster rate than for all included single target covariances.

Theorem 4.5 shows that MTS of the covariance is consistent in the LDL. It can also be seen that data sets with covariance distance (C1) increasing faster than $O(p)$ do not contribute to the MTS estimator in the LDL limit: for $n \to \infty$, sample covariances $\tilde{C}^k$ of these data sets do not constitute useful shrinkage targets because the limit behaviour of the sample covariance $\tilde{C}$ is superior.

**FOLDL consistency of MTS of the covariance** Let us now consider the case where only the dimensionality $p$ goes to infinity, while $n$ remains constant.

**Theorem 4.6** (FOLDL consistency of MTS of the covariance). Let us assume a sequence of statistical models indexed by $p$ for which (D1), (D2), (D3), (D4), (C1), (C2) (see Theorem 4.5) and

\[(C3') \quad \tau^C_k < 1 \quad \text{and} \quad \tau_\gamma < 1\]

\[(C5) \quad \sum_{i,j,k,l \in Q_p} \text{Cov} \left[ (y_{i1}y_{j1})^2, (y_{k1}y_{l1})^2 \right] \frac{1}{|Q_p|} = o(1)\]

hold. Then, for the set of targets in (Schäfer & Strimmer, 2005) and targets given by additional data sets, assumptions (G1), (G2), (G3), and (G4) of Theorem 4.2 are fulfilled, and MTS of the covariance and $\tilde{\Lambda}$ are consistent.

Proof. see appendix D.6.

As for the mean, consistency in the FOLDL requires a restriction (C3') on the largest eigenvalue (compare to Theorem 4.4). Assumption (C5) further restricts covariances between uncorrelated random variables. Note that identifiability holds even without assumption (C4).

### 4.6 Simulations

The proposed MTS has more free parameters than STS (LW-Shrinkage) and therefore $\tilde{\lambda}$, the vector of shrinkage intensity estimates, has a higher variance than the single shrinkage intensity estimate $\hat{\lambda}$ in STS.

In this section, simulations for both MTS of the mean and MTS of the covariance are provided which show that already at moderate data set sizes, MTS accurately estimates $\lambda$.

MTS is evaluated with respect to both estimation error, measured by PRIAL, and performance, measured by classification accuracy. MTS of the mean is evaluated in an application of Linear Discriminant Analysis (LDA), MTS of the covariance in Common Spatial Patterns as a preprocessing for LDA.

#### 4.6.1 Simulations for MTS of the mean

**Simulation 1: MTS of the mean to additional data sets**

In the first simulation, the behavior of MTS of the mean in the large-dimensional limit (LDL, $p, n \to \infty$) is illustrated. Artificial data with $n$ standard normal data points of dimensionality $p = n$ with mean
\[ \mu_i = 0 \] is generated. For the shrinkage targets, \( K = 4 \) standard normal data sets are generated, with \( n^k = p \) data points and means \( \mu^k = (\pm 1)_i \eta_k \), where the sign is random and \( \eta = (\sqrt{p}^{-1}, 0.5, 1.0, 2.0)/5 \) defines the quality of the four additional data sets. In this setting, the first additional data set \( X^1 \) has \( \tau^1_p = 0 \) and \( X^{2/3/4} \) have \( \tau^{2/3/4}_p = 1 \).

This setting fulfills the assumptions of Theorem 4.3: targets have a clear limit behavior (M1), from standard normality follows \( \tau^{(k)} = 0 \) (M2) and the means of the targets are independently sampled (M3). The theorem tells us that the MTS estimator converges and that targets \( T^{2/3/4} \) do not receive any weight in the LDL.

MTS is compared to five versions of STS: STS to each of the targets \( \hat{T}^k = \hat{\mu}^k \) and STS to the joint target \( \hat{T}^{\text{joint}} := \hat{\mu}^{\text{joint}} := 0.25 \cdot \sum_k \hat{\mu}^k \). Figure 4.4 shows the dependency of the PRIAL (left) and the shrinkage intensities (right) on the dimensionality \( p \).

As predicted for the LDL by Theorem 4.3, the STS and MTS shrinkage intensities for targets \( \hat{\mu}^2, \hat{\mu}^3, \hat{\mu}^4 \) and \( \hat{\mu}^{\text{joint}} \) go to zero: these targets are not useful in the limit. Only the target \( \hat{\mu}^1 \) remains useful. As \( n = n_1 \) and the entries is \( \mu^1 \) converge to the entries in \( \mu \), the shrinkage intensity \( \lambda^1 \) goes to 0.5.

The PRIALs reflect this picture: For the asymptotically useless targets, the improvement over the sample mean goes to zero, for \( \hat{\mu}^1 \) it goes to a constant. For low \( p \) and \( n \), it is less relevant that \( \mu^2, \mu^3 \) and \( \mu^4 \) are different from \( \mu \); as a consequence, the joint target is better than \( \mu^1 \). Over the whole range of \( p \), \( \hat{\mu}^{\text{MTS}} \) outperforms all STS estimators. For \( p \rightarrow \infty \), MTS converges to \( \hat{\mu}^1 \). Figure 4.5 shows convergence for the finite observations large-dimensional limit (FOLDL). The experiment is analogous to the one above, only \( n = n^k = 50 \) is kept fixed. Contrary to the LDL, all shrinkage intensities remain finite. As above, over the whole range of \( p \), \( \hat{\mu}^{\text{MTS}} \) outperforms all STS estimators.

**Simulation 2: MTS for Linear Discriminant Analysis**

To test MTS in a classification setting, the simulation above is extended to two class means \( \mu_{A/B} \) (\( p = 50 \), \( n = 50 \)). The difference of the class means is identical in each dimension, chosen such that the Bayes optimal classifier achieves 80% accuracy. For both classes there are four additional data sets, \( n^k = 100 \) with mean differences

\[ \Delta \mu^k_{A/B,i} = \mu^k_{A/B,i} - \mu_{A/B,i} = (\pm 1)_i \eta_k, \]

\[ \|\mu - \mu^1\| > \|\mu - \mu^2\| \]
4.6. SIMULATIONS

Figure 4.5. Finite observations large-dimensional limit (FOLDL) of MTS of the mean to additional data sets. Average obtained over $R_r = 20$ repetitions for $R_m = 500$ models. Shaded areas show one standard deviation.

$\eta = 10^k \cdot (0.25, 0.5, 1, 2)$ where the parameter $\kappa$ governs the similarity of the additional data sets. The covariance of each data set is $C_{A/B}^{(k)} = \mathbf{I}$. To make the setting slightly more realistic, the data set is transformed to have diagonal covariance with eigenvalues $\gamma_i = 10^{(i-1)/(p-1)}$ (log-spaced between $10^{\pm \alpha}$, $\alpha = 1$). This is achieved by rescaling all data points:

$$x^{(k)}_{A/B, it} \leftarrow x^{(k)}_{A/B, it} \cdot \sqrt{\gamma_i}.$$

Linear Discriminant Analysis using different mean estimators is applied: MTS is compared to (A) sample means $\hat{\mu}_{A/B}$, where the additional data sets are ignored\(^8\), (B) pooled means where $\hat{\mu}_{A/B}^{\text{pooled}} := (K+1)^{-1}(\sum_k \hat{\mu}_{A/B}^k + \hat{\mu}_{A/B})$, and (C) STS where both sample means $\hat{\mu}_{A/B}$ are shrunk to the corresponding joint target $\hat{\mu}_{A/B}^{\text{joint}} := K^{-1} \sum_k \hat{\mu}_{A/B}^k$.

Figure 4.6 (left) shows the gain in classification accuracy relative to the baseline of sample means in dependence of the scale parameter $\kappa$. When the target means are very similar ($\kappa \to -\infty$), pooled means is the optimal solution. For very different target distributions ($\kappa \to \infty$) it is not possible to improve over the sample means $\hat{\mu}_{A/B}$. For these extreme cases, STS to the pooled data performs as well as pooled means and sample means, respectively. For intermediate values of $\kappa$, it outperforms both approaches. MTS improves on STS by finding a superior weighting of the target means.

For Figure 4.6 (right), a spike has been added to the covariance model: The largest eigenvalue has been multiplied by 100 and the corresponding direction has been made non-discriminative. The drop in performance indicates that now the ESE is no longer a good proxy for classification accuracies: STS and MTS now give too much weight to the targets, especially to the less useful targets $\hat{\mu}_{A/B}^{3/4}$. All targets are similar to the original data in the non-discriminative direction of the spike, but still vary in quality in the discriminative directions.

**Whitening — a practical trick** Shrinkage puts too much weight on the direction of highest variance. Whitening the data before MTS (wMTS) helps: wMTS gives equal importance to all directions, yields proper weights for the $\hat{\mu}_{A/B}^k$ and superior accuracies.

Interestingly, wMTS also performs better than standard MTS when there is no spike in the covariance (left). In this case the estimation of the shrinkage intensities is dominated by the few directions of largest variance. This causes high variance in the shrinkage intensity estimates $\tilde{\lambda}$. Using wMTS, the

---

\(^8\)To increase comparability, the sample covariance is averaged over all data sets, independently of the estimator of the mean.
CHAPTER 4. REGULARIZED ESTIMATION II: MTS

Figure 4.6. Accuracy gain for MTS for Linear Discriminant Analysis. Average obtained over \( R_r = 20 \) repetitions for \( R_m = 500 \) models. Shaded areas show one fourth standard deviation.

Figure 4.7. Large-dimensional limit (LDL) of MTS of the covariance to additional data sets. Average obtained over \( R_r = 20 \) repetitions for \( R_m = 500 \) models. Shaded areas show one standard deviation.

estimation of the shrinkage intensities becomes an evenly weighted average over dimensions and hence gets more stable.

In general, whitening leads to large improvements if the discriminative information is not restricted to the subspace of highest variance.

4.6.2 Simulations for MTS of the covariance

Simulation 3: MTS of the covariance to additional data sets

Here, the behavior of MTS of the covariance in the large-dimensional limit (LDL, \( p, n \to \infty \)) is illustrated. Artificial data is generated with \( n \) normal data points of dimensionality \( p = n \), their covariance \( \mathbf{C} \) is diagonal with logarithmically spaced eigenvalues. For the shrinkage targets, \( K = 4 \) normal data sets with \( n^k = p \) data points are generated. The covariance matrices \( \mathbf{C}^k \) of the additional data sets only differ in the largest eigenvalue \( \lambda_{\max}^k = \eta_k \cdot p \), with \( \eta = (\sqrt{p}^{-1}, 1.0, 2.5, 5.0)/10 \). Therefore the first additional data set \( \mathbf{X}^1 \) has \( \tau^1_C = 1 \) and \( \mathbf{X}^{2/3/4} \) have \( \tau^1_C = 2 \).

This makes the setting analogous to Simulation 1. Figure 4.7 shows the dependency of the PRIAL (left) and the shrinkage intensities (right) on the dimensionality \( p \): the STS and MTS shrinkage intensities for targets \( \mathbf{C}^{2/3/4} \) and \( \mathbf{C}^{\text{joint}} \) go to zero, only the target \( \mathbf{C}^1 \) remains useful in the LDL. As \( n = n^1 \), the
4.6. SIMULATIONS

Figure 4.8. Finite observations large-dimensional limit (FOLDL) of MTS of the covariance to additional data sets. Average obtained over $R_r = 20$ repetitions for $R_m = 500$ models. Shaded areas show one standard deviation.

Figure 4.9. MTS of the covariance to identity and additional data sets. Average obtained over $R_r = 20$ repetitions for $R_m = 500$ models. Shaded areas show four standard deviations.

shrinkage intensity goes to 0.5. For the asymptotically useless targets, the PRIAL over the sample covariance goes to zero, for $\mathbf{C}^1$ it goes to a constant. For low $p$ and $n$, it is less relevant that $\mathbf{C}^{2/3/4}$ are different from $\mathbf{C}$: as a consequence, the joint target is better than $\mathbf{C}^1$. Over the whole range of $p$, $\hat{\mathbf{C}}^{\text{MTS}}$ outperforms all STS estimators.

Figure 4.8 shows results for the FOLDL, where $n = n^1 = n^2 = n^3 = n^4 = 50$ is kept fixed. As for the mean, all shrinkage intensities remain finite and over the whole range of $p$, $\hat{\mathbf{C}}^{\text{MTS}}$ outperforms all STS estimators.

Simulation 4: MTS with identity target and additional data sets

For MTS of the covariance there is also the possibility to include a biased estimator as a shrinkage target. The most widely used biased estimator is the identity multiplied by the average sample eigenvalue: $\mathbf{T}^{\text{id}} := \frac{1}{n} \mathbf{I}$. In this simulation, the sample covariance is shrunk to $\mathbf{T}^{\text{id}}$ and the covariance matrices of four additional sets of observations. $\mathbf{C}$ and $\mathbf{C}^\text{id}$ are chosen diagonal with logarithmically spaced eigenvalues between $10^{-1}$ and $10^1$. Five multivariate normal random data sets $\mathbf{X}$ and $\mathbf{X}^{1/2/3/4}$, of dimensionality $p = 500$, are generated with $n = 500$, $n^1 = p/2$, $n^2 = p$, $n^3 = 2p$ and $n^4 = 4p$ observations. Each of the additional data sets is rotated randomly constrained to a rotation angle $\phi$. 
Figure 4.10 shows PRIAL and shrinkage intensities in dependence of the rotation angle $\phi$. Shrinkage to $T^d$ is independent of $\phi$, while STS to the other data is good when distributions are similar (small rotation angle) and yields only small improvements for very different distributions (large rotation angle). The MTS shrinkage intensities show that for large $\phi$ MTS yields approximately the same estimate as STS to $T^d$, while for small $\phi$ it yields a weighting of all five targets. This weighting yields superior PRIAL compared to each STS estimator.

Simulation 5: MTS of the covariance and CSP

In this subsection, MTS is applied to the preprocessing method Common Spatial Patterns (CSP, see Section 3.3.4).

For the simulation, a diagonal covariance matrix $C$ of dimensionality $p = 50$, with logarithmically spaced eigenvalues between $10^{-1}$ and $10^{1}$, is generated. The covariances of the two classes $C^{A,B}$ and a set of different covariances $C^{A/B,k}$ are each obtained by rescaling $P = 10$ random eigenvalues of $C$ by $p_i = (1 + i/P), i = 1,2,\ldots,P$. In addition, the $C^{A/B,k}$ are rotated randomly by an angle $\phi^k$, $\phi = (0,5,10,90)$. To study the dependency on the similarity of targets, the covariance matrices of the additional data sets are set to

$$C^{A,B,k}(w) = (1-w)C_{\text{diff}}^{A/B,k} + wC^{A/B}.$$ 

For each class and each target, $n = n^k = 200$ data points are generated. The classification accuracy is calculated for test trials of length $n^{\text{test}} = 20$.

Figure 4.10 (left) shows the relative classification accuracies of the different covariance estimation approaches. For $w = 1$, the target covariances are equal to the class covariances and $S^{\text{pooled}} = 1/(k+1)(\sum_k S^k + S)$ is optimal. For $w \to 0$, the targets do not contain discriminative information, hence the sample covariance becomes optimal. STS to the joint covariance of the additional data sets performs better than the pooled covariance, but is clearly outperformed by MTS. Whitened MTS performs even better.

For Figure 4.10 (right) a spike has been added to all covariance matrices: The largest eigenvalue has been multiplied by 100 and the corresponding direction was excluded from the random rotations. This strong direction dominates the standard STS and MTS estimates and causes a strong degradation of performance: the ESE is no longer a good proxy for classification accuracy. The performance of whitened MTS, on the other hand, is not affected.
4.7 Multi-Target Shrinkage on real world data

This section spotlights two application scenarios of MTS on real world data, one for MTS of the mean and one for MTS of the covariance.

4.7.1 MTS of the mean for ERP-BCI

MTS of the mean is applied to the ERP-BCI data set from Schreuder et al. (2011) described in Section 1.6.2. Normally, LDA classifiers for ERP-BCI are computed by pooling all target and all non-target data, respectively, thus neglecting the stimulus identity. Note that despite having the same name, there is no relation between the targets in an ERP experiment and shrinkage targets.

In the following, three alternatives which yield a binary classifier for each stimulus are analyzed. The stimulus-specific estimators require estimates of the target and the non-target mean for each stimulus. Applying STS and MTS, the stimulus-specific means are calculated using the mean over the distinct stimulus classes as a shrinkage target (STS) or each mean of each distinct stimulus class as a separate shrinkage target (MTS). The stimulus-specific sample means are included as a baseline comparison.

In ERP analysis, the covariance can be considered as general background activity which is independent of the stimulus. Hence, for all approaches the pooled covariance is used. As the spectrum in the ERP-BCI data is heavily tilted, whitening is applied as a MTS preprocessing.

Figure 4.11 shows the classification accuracies based on wMTS estimation of subclass means, comparing against classification accuracies based on pooled means (standard approach), (subclass-specific) sample means and wSTS estimation of subclass means. The analysis shows the wMTS estimation of subclass means to be superior to all other approaches.

In ERP analysis, the covariance can be considered as general background activity which is independent of the stimulus. Hence, for all approaches the pooled covariance is used. As the spectrum in the ERP-BCI data is heavily tilted, whitening is applied as a MTS preprocessing.

There exist extensions of LDA which consider subclass labels by fitting a global classifier based on Gaussian mixtures (Hastie & Tibshirani, 1996; Gkalelis et al., 2011). The advantage of the approach presented here is that it yields an optimized classifier for each subclass.

4.7.2 MTS of the covariance matrix for MI-BCI

To test MTS of the covariance, the MI-BCI data set from Blankertz et al. (2010a), which was described in Section 1.6.2 and already analyzed in Section 3.3.4, is reanalyzed. As training is expensive, exploiting training data from other subjects is desirable. Two approaches are compared: STS to the covariance of all other subjects and Multi-Target Shrinkage to the covariances of all 80 subjects.
Directions of high variance dominate shrinkage estimators (Bartz & Müller, 2013) and the BCI data contains pronounced directions of high variance, the spectrum is heavily tilted. To reduce the impact of the first eigendirections without giving too much importance to low-variance noise directions, a special form of whitening is applied: only for the calculation of the shrinkage intensities, the first five principal components were rescaled to have the same variance as the sixth principal component. Shrinkage is corrected for autocorrelation, as described in Section 3.3.

Figure 4.12 (left, middle) shows accuracies for various numbers of training trials per class. It can be seen that STS outperforms sample covariance matrices, while it is not possible to estimate the high number of parameters for MTS. For few training trials, wSTS outperforms STS, as the averaging over additional dimensions reduces variance. wMTS yields very good accuracies.

Figure 4.12 (right) shows shrinkage intensities. It can be seen why MTS fails: when shrinkage is dominated by the first eigendirections, there are targets which appear too good and receive very large shrinkage intensities. Whitened MTS takes more directions into account and yields lower shrinkage intensities.

Figure 4.13 shows subject-wise accuracies for the different covariance matrix approaches for ten training trials. The proposed wMTS estimator significantly outperforms all other approaches.

4.8 Chapter Summary

In the last chapter, LW-Shrinkage was presented as a widely applied regularization technique for covariance estimation. In this chapter, I pointed out several use cases in which a single shrinkage target is not...
sufficient. This motivates the usage of multiple shrinkage targets.

I derived estimator-independent formulas for optimal Multi-Target Shrinkage (MTS) and theoretically analyzed the necessary conditions for consistency. For the specific instances of sample mean and sample covariance, I showed when these conditions are fulfilled and provided simulations which demonstrate that MTS yields improvements over STS in several situations. As a practical trick, I proposed whitening as a preprocessing step which increases the robustness of MTS.

For both MTS of the mean and MTS of the covariance, I provided a real world example from the neuroscience domain. A detailed analysis showed that the proposed method yields a significant performance enhancement over Single-Target Shrinkage.
Chapter 5

Spectrum Correction

The previous sections dealt with shrinkage techniques for covariance estimation, the most popular being LW-Shrinkage. These methods are governed by a single regularization parameter, which makes them fast, easy to implement and numerically robust. The downside is that this limits their flexibility. In general, shrinkage is not flexible enough to improve estimation in the low- and high-variance regions of the spectrum at the same time.

In this chapter, approaches which correct each eigenvalue separately are discussed. Although Charles Stein already proposed the first algorithm in 1975, spectrum correction has found little application in practice: in these algorithms, the estimation of the large number of correction factors suffers from a high variance.

In the past few years, researchers have increasingly begun work on the Random Matrix Theory (RMT) of covariance matrix estimation. The standard problem of RMT is the derivation of the distribution of the sample eigenvalues from the distribution of the population eigenvalues. The relationship is governed by the Marčenko-Pastur law. In covariance matrix estimation the direction of inference is inverted: the sample eigenvalues are given and we are interested in the population covariance matrix. El Karoui (2008b) and Ledoit & Péché (2011) discuss covariance estimation by inversion of the Marčenko-Pastur law, research which led to the state-of-the-art approach Nonlinear Shrinkage (NLS).

In this chapter, I propose CVC, a cross-validation-based spectrum correction approach which estimates the corrected eigenvalues on hold-out sets. The CV-approach optimizes the same distance measure as NLS and has the advantage of conceptual simplicity, smaller runtime and higher numerical stability.

Theoretical guarantees for cross-validation require hypothesis or error stability of the algorithm under consideration (Vapnik & Kotz, 1982; Devroye et al., 2013; Kearns & Ron, 1999). The algorithmic instability of the eigenvalue decomposition, which I illustrate in detail, makes it very hard to obtain theoretical guarantees for CVC, hence this left for future work. This lack of a theoretical foundation is compensated by very strong empirical results: the evaluation of CVC on real world data sets from four domains shows that CVC is the best performing general purpose covariance estimator. The chapter is based on an article in preparation (Bartz, 2015a).

Chapter outline An introduction to spectrum correction and its history is given in Section 5.1. Section 5.2 discusses approaches based on Random Matrix Theory: the inversion of the direction of inference in RMT is discussed in Section 5.2.1 and Section 5.2.2 presents Nonlinear Shrinkage, the state-of-the-art approach for spectrum correction. Then, in Section 5.3, CVC is proposed, the CV-based spectrum correction approach. Performance on real world data is shown in Section 5.4.
CHAPTER 5. SPECTRUM CORRECTION

5.1 Improved estimation by spectrum correction

The sample covariance matrix is a symmetric matrix. It can be decomposed according to

\[ \mathbf{S} = \mathbf{VV}^T = \sum_i \hat{\gamma}_i \hat{\mathbf{v}}_i \hat{\mathbf{v}}_i^T, \]

where \( \hat{\mathbf{V}} \) is a rotation matrix given by the eigenvectors \( \hat{\mathbf{v}}_i \) and \( \hat{\mathbf{F}} \) is the diagonal matrix of sample eigenvalues \( \hat{\gamma}_i \). Shrinkage to the standard target (eq. (3.1) and eq. (3.2)) can also be seen as an approach which retains the eigenbasis and modifies the eigenvalues:

\[ \hat{\gamma}_i^{\text{shr}} := (1 - \lambda) \hat{\gamma}_i + \lambda p^{-1} \sum_i \hat{\gamma}_i, \]

(5.1)

First approaches to spectrum correction The convex combination in eq. (5.1) is governed by a single parameter. More flexible approaches modify each eigenvalue individually. The first one was proposed by Stein in his 1975 Rietz lecture\(^1\),

\[ \hat{\gamma}_i^{\text{Stein}} := \frac{\hat{\gamma}_i}{n - p + 1 + \hat{\gamma}_i \sum_{j \neq i} (\hat{\gamma}_i - \hat{\gamma}_j)^{-1}}, \]

(5.2)

and minimizes an approximation of Stein’s Loss. Stein (1986) later proposed a slightly more complex extension of this estimator.

The obtained eigenvalues may no longer decrease monotonically. Hence, it is recommended to apply isotonic regression which yields a monotonically decreasing sequence (for details, see Section 5.3.2)

Dey & Srinivasan (1985) proposed a set of alternative spectrum correction methods based on the minimax principle, the simplest is given by

\[ \hat{\gamma}_i^{\text{Dey}} := \frac{n}{n + p + 1} \hat{\gamma}_i. \]

(5.3)

Spectrum correction has found little application in practice. Although the estimation of \( p \) correction factors is very powerful, the higher flexibility comes at the cost of higher estimation errors. Both approaches presented in this paragraph tend to perform worse than the optimal regularization of the whole covariance matrix with a single regularization parameter (Ledoit & Wolf, 2004). In addition, the reciprocal in \( \hat{\gamma}_i^{\text{Stein}} \) is numerically unstable and \( \hat{\gamma}_i^{\text{Dey}} \) is not defined for \( p > n \).

5.2 Random Matrix Theory (RMT) for spectrum correction

The field of RMT studies the theoretical properties of large random matrices (for an overview, see Edelman & Rao (2005)). The majority of the research focuses on the task of inferring the distribution of the sample eigenvalues from the distribution of the entries in a random matrix. For covariance matrices, this distribution is given by the Marchenko-Pastur law and its generalizations (Marchenko & Pastur, 1967; Silverstein, 1995).

Consequently, the first applications of RMT to covariance estimation only used the sample eigenvalue distribution, to estimate the number of factors in a factor model. Rosenow et al. (2002) set the number of factors to the number of sample eigenvalues larger than the largest eigenvalue of a random matrix, assuming that the corresponding eigendirections reflect some real structure. Their model is equivalent to a PCA factor model based on the correlation matrix. Laloux et al. (2000) propose a very similar approach based on a probabilistic PCA model.\(^2\)

---

\(^1\)As there is no corresponding publication, the presentation here has to rely on the presentation in (Dey & Srinivasan, 1985).

\(^2\)As there is no corresponding publication, the presentation here has to rely on the presentation in (Dey & Srinivasan, 1985).
5.2. RANDOM MATRIX THEORY (RMT) FOR SPECTRUM CORRECTION

5.2.1 Inverting the Marčenko-Pastur law

El Karoui (2008b) made the first real attempt to harness the possibilities of RMT. In his work, he numerically inverts the Marčenko-Pastur law, which describes the relationship between the distributions of population and sample eigenvalues, in order to estimate the population eigenvalues from the sample eigenvalues. The following paragraphs give a very short introduction to this approach.

Stieltjes transforms and the MP-law

RMT is naturally formulated in terms of distributions: the population eigenvalues are not fixed, but drawn from a distribution. The distribution of the sample eigenvalues then depends on the distribution of population eigenvalues.

El Karoui (2008b) considers asymptotics at fixed spectral distribution: the cumulative distribution function (cdf) of the population eigenvalues $H_p$ is kept constant: $H_p = H_\infty, \forall p$. Then, the empirical cdf $\hat{F}_p$ of the sample eigenvalues is asymptotically non-random and the relation between $H_\infty$ and $F_\infty$ is given by the Marčenko-Pastur law.

Many derivations and proofs in RMT rely on the Stieltjes transform of the distribution of sample eigenvalues. This often simplifies the theoretical analysis. The Stieltjes transform of the cdf $bF_p$ on $\mathbb{R}$ is a function with complex arguments given by

$$m_{bF_p}(z) := \int \frac{dF_p(x)}{x-z}, \forall z \in \mathbb{C}^+, \quad (5.4)$$

where $\mathbb{C}^+ := \{z \in \mathbb{C} : \text{Im}(z) > 0\}$.

El Karoui (2008b) analyzes $\hat{G}_p$, the empirical spectral distribution of the $n \times n$-matrix $n^{-1}X^TX$. In Machine Learning, this is called the linear kernel matrix (Schölkopf & Smola, 2002). Its eigenvalues are, excluding zero eigenvalues, equal to the sample eigenvalues.

Silverstein (1995) derived the most general form of the Marčenko-Pastur law, which El Karoui (2008b) uses to relate (the Stieltjes transform of) $\hat{G}_p$ to the unknown distribution of population eigenvalues $H_\infty$:

$$m_{G_\infty}(z) = -\left(z - c \int \frac{\gamma dH_\infty(\gamma)}{1 + \gamma m_{G_\infty}(z)}\right)^{-1}. \quad (5.5)$$

Note that the Stieltjes transform $m_{G_\infty}(z)$ appears on both sides of the equation.

Solving for the distribution of population eigenvalues

To estimate the population eigenvalues, one has to find the $H_\infty$ which best satisfies Equation (5.5) for $m_{\hat{G}_p}$, the Stieltjes transform of $\hat{G}_p$. To make this feasible, El Karoui (2008b) makes two approximations:

1. He approximates $dH_\infty(\gamma)$, the pdf of population eigenvalues, by a sum of delta peaks:

$$dH_\infty(\gamma) \approx \sum_{k=1}^K w_k \delta(t_k - \gamma), \quad w_k \geq 0 \text{ and } \sum_k w_k = 1, \quad (5.6)$$

where the $t_k$ form a predefined grid. It is possible to use different or include additional basis functions.

2. Instead of optimizing eq. (5.5) for all $z \in \mathbb{C}^+$, he only considers a discrete set of $J$ points $(z_j, m_{G_\infty}(z_j))$.

Plugging the approximation eq. (5.6) into the MP-law eq. (5.5) leads to a set of $J$ complex errors

$$e_j := \frac{1}{m_{\hat{G}_p}(z_j)} + z_j + \frac{p}{n} \sum_{k=1}^K w_k \frac{t_k}{1 + m_{\hat{G}_p}(z_j)t_k},$$

which can be minimized by searching over the weights $w_k$. In the next step, the $p$ population eigenvalues are obtained by calculating the quantiles of $dH_\infty(\gamma)$. 


**Challenges**  Unfortunately, El Karoui’s approach has a set of drawbacks:

- The largest eigenvalue of the covariance matrix is often isolated from the bulk. The discrete approximation of the population spectrum makes this case problematic.
- It is not clear how the choices of the set \((z_j, m_G(z_j))\) and the grid \(t_k\) affect the algorithm.
- The description of the algorithm is not clear. Ledoit et al. (2012) state that they were not able to reproduce the results and that the original code is not available.
- For spectrum correction, the population eigenvalues are not optimal w.r.t. the ESE (see Section 3.3), because the sample eigenvectors differ from the population eigenvectors. Instead, the optimal corrected eigenvalues are given by

\[
\gamma^* := \arg \min_{\gamma} \left\| C - \sum_i \gamma_i \tilde{v}_i \tilde{v}_i^\top \right\|^2 \quad \iff \quad \gamma^*_i = \tilde{v}_i^\top C \tilde{v}_i. \tag{5.7}
\]

In the following, \(\gamma^*\) is called the (spectrum) oracle.

These aspects make it difficult to apply the El Karoui approach in practice. To my knowledge, no publication reports a successful application to a real world data set.

### 5.2.2 Adapting RMT to the covariance estimation problem

In the last subsection, it was mentioned that estimating the population eigenvalues is not optimal for spectrum correction. Recently, Ledoit & Péché (2011) quantified the relationship between sample and population eigenvectors. Based on this relationship, they derived the distribution of the variances in the direction of the sample eigenvectors (see eq. (5.7)). This directly yields an estimator for the spectrum oracle \(\gamma^*\) which Ledoit et al. (2012) call **Nonlinear Shrinkage (NLS)**:

\[
\gamma^* \text{NLS}_i := \hat{\gamma}_i \left| 1 - \frac{1}{c} - e^{-1} \hat{\gamma}_i \hat{m}_{F_\infty}(\hat{\gamma}_i) \right|^{-2}, \tag{5.8}
\]

where \(\hat{m}_{F_\infty}(z)\) is an estimate of the Stieltjes transform of the limit distribution of the sample eigenvalues\(^2\).

Ledoit et al. (2012) obtain \(\hat{m}_{F_\infty}(z)\) from the Marčenko-Pastur law and an estimate of the population eigenvalues. They propose an alternative to the inversion procedure presented in the last section which directly optimizes for the population eigenvalues.

The simulations in the next section use a different version which is more stable\(^3\) (Ledoit & Wolf, 2014). It is based on an alternative formulation of eq. (5.5) in (Silverstein, 1995): for the limit distribution of sample eigenvalues \(F_\infty(z)\), the MP-law is given by

\[
m_{F_\infty}(z) = \int \frac{dH_\infty(\gamma)}{\gamma(1 - e^{-1} - e^{-1} z \hat{m}_{F_\infty}(\gamma)) - z}. \tag{5.9}
\]

Ledoit & Wolf (2014) discretize eq. (5.9) and derive a function \(Q : [0, \infty)^p \to [0, \infty)^p\) which maps population eigenvalues to sample eigenvalues. The function \(Q\) is used to formulate a (non-convex) optimization problem for the population eigenvalues,

\[
\hat{\gamma}^{pop} := \arg \min_{\gamma} \| Q(\gamma) - \hat{\gamma} \|^2,
\]

which can be solved by Sequential Linear Programming (SLP).

---

\(^2\)The Stieltjes transform \(\hat{m}_{F_\infty}(z)\) is only defined for \(z\) with positive imaginary parts (see eq. (5.4)). For real values, one defines \(\hat{m}_{F_\infty}(\hat{\gamma}_i) := \lim_{y \to 0^+} \hat{m}_{F_\infty}(\gamma_i + iy)\).

\(^3\)Personal correspondence with Olivier Ledoit.
5.3 Using cross-validation for covariance estimation

Ledoit et al. (2012) state that anonymous reviewers proposed to use leave-one-out cross-validation for the estimation of the variances in eq. (5.7). The leave-one-out (loo) cross-validation covariance (CVC) estimator is defined by

\[
\hat{\gamma}^{\text{loo-CVC}}_i := \frac{1}{n} \sum_{t=1}^n \left( \hat{\mathbf{v}}^{\text{loo}(t)}_i \mathbf{x}_t \right)^2,
\]

(5.10)

where \(\hat{\mathbf{v}}^{\text{loo}(t)}_i\) is the \(i\)th sample eigenvector of the sample covariance \(\hat{\mathbf{S}}^{\text{loo}(t)}\), for which the \(i\)th data point has been removed from the data set. The vectors \(\hat{\mathbf{v}}^{\text{loo}(t)}_i\) and \(\mathbf{x}_t\) are independent, this makes \(\hat{\gamma}^{\text{loo-CVC}}_i\) an unbiased estimator\(^4\) of the variance in the directions \(\hat{\mathbf{v}}^{\text{loo}(t)}_i\):

\[
\mathbb{E} \left[ \hat{\gamma}^{\text{loo-CVC}}_i \right] = \frac{1}{n} \sum_{t=1}^n \mathbb{E} \left[ \hat{\mathbf{v}}^{\text{loo}(t)}_i \mathbf{x}_t \hat{\mathbf{v}}^{\text{loo}(t)}_i \right] = \mathbb{E} \left[ \hat{\mathbf{v}}^{\text{loo}(t)}_i \mathbf{C} \hat{\mathbf{v}}^{\text{loo}(t)}_i \right].
\]

(5.11)

Ledoit et al. (2012) state that the CVC approach has the advantage of conceptual simplicity, while NLS has three clear advantages:

1. NLS can estimate any smooth function of the eigenvalues. In particular, it is possible to directly optimize for the precision matrix.

2. NLS yields clearly smaller estimation error than CVC.

3. NLS is faster than the CVC approach. CVC requires \(n + 1\) eigendecompositions of \(p\)-dimensional sample covariance matrices. For large \(p\), this is very time consuming.

The following subsections discuss these aspects.

5.3.1 Precision matrix estimation

An intriguing advantage of the NLS technique is the possibility of estimating the precision matrix directly. Interestingly, the optimal spectrum correction is different for covariance and precision matrix:\(^5\):

\[
\gamma^\circ := \arg \min_{\gamma} \left\| \mathbf{C}^{-1} - \sum_i \gamma_i^{-1} \mathbf{v}_i \mathbf{v}_i^\top \right\|^2 \quad \iff \quad \gamma_i^\circ^{-1} = \mathbf{v}_i^\top \mathbf{C}^{-1} \mathbf{v}_i \neq \gamma_i^{-1}.
\]

(5.12)

In the following, \(\gamma^\circ\) is called the precision oracle. Based on the results on the relationship between sample and population eigenvectors, Ledoit & Péché (2011) derive an estimator for \(\gamma^\circ\):

\[
\hat{\gamma}_i^{\text{NLS-precision}} := \hat{\gamma}_i \left( 1 - c^{-1} - 2c^{-1}\hat{\gamma}_i \text{Re}[\hat{m}_{F^\circ}(\hat{\gamma}_i)] \right)^{-1}.
\]

Choice of loss function  Many algorithms require estimates of the precision matrix and hence it seems to be a good idea to use a covariance estimator with minimal ESE for the precision. One has to keep in mind that in a practical application, both definitions of estimation error only serve as proxies for the performance. It is not possible to say in general that one of the two will work better. Two aspects have to be kept in mind:

- The choice between \(\hat{\gamma}^\ast\), an estimator with lower ESE for the covariance, and \(\hat{\gamma}^\circ\), an estimator with lower ESE for the precision, is not obvious. Consider portfolio optimization:

\(^4\)Note that the normalization by \(n\) only yields an unbiased estimate because the data is assumed to have zero mean.

\(^5\)Ledoit & Wolf (2013) also discuss optimal estimation under Stein’s Loss.
On the one hand, using $\hat{\gamma}^\circ$ yields smaller estimation errors in the portfolio weights than $\hat{\gamma}^\ast$. On the other hand, $\hat{\gamma}^\ast$ yields less estimation error in the portfolio variance than $\hat{\gamma}^\circ$. As a consequence, it is possible that the small estimation error in the weights obtained using $\hat{\gamma}^\circ$ lies in directions which incur higher portfolio risk than the error in the weights obtained using $\hat{\gamma}^\ast$.

The main advantage of spectrum correction compared to LW-Shrinkage, the individual optimization of eigenvalues, holds independently of the specific loss function: spectrum correction is flexible enough to precisely estimate low- and high-variance directions, while in LW-Shrinkage, strong directions dominate the loss function and large relative estimation errors in weak directions are neglected.

Appendix A provides simulations which compare LDA classification performances for $\hat{\gamma}^\ast$ and $\hat{\gamma}^\circ$. The results support the hypothesis that most applications will not display a pronounced performance difference between both estimators.

5.3.2 Estimation error of the CVC approach

Ledoit et al. (2012) show that loo-CVC does not yield competitive results. In the following paragraphs, I (i) explain what makes a theoretical analysis difficult, (ii) illustrate why loo-CVC is not competitive and (iii) propose iso-loo-CVC, an improved CVC estimator.

Theoretical analysis of loo-CVC The loo-CVC estimator yields unbiased estimates of the variance in the eigendirections for $n - 1$ data points, as shown in eq. (5.11), and is therefore asymptotically unbiased. In general, convergence and the convergence rate of cross validation depend on the algorithm under consideration (Vapnik & Kotz, 1982; Devroye et al., 2013) and require some notion of algorithmic stability (Kearns & Ron, 1999): removing a single data point from the data set is assumed to have negligible influence on the obtained solution (hypothesis stability) or out-of-sample-error (error stability). The eigendecomposition is not algorithmically stable: the removal of a single data point can completely change the obtained eigendirections. In fact, the dependency between the loo-eigendirections $\tilde{b}_v^{loo}(i)$ causes a high estimation error in the loo-estimates of the variances (eq. (5.10)). A detailed theoretical analysis is left for future research.

Illustration of the loo-CVC estimation error Figure 5.1 illustrates the high estimation error of loo-CVC for Gaussian data with an identity population covariance ($p = 50$, $n = 1000$). For identity covariance, the loo-sample eigenvectors are random projections which are dependent over folds. To show the effect of this dependence, loo-sample eigenvectors are compared to projections which are entirely random.

The upper left plot shows (estimated) variances given by population and sample eigenvalues as well as estimates based on loo-CVC and random projection. It can be seen that the variance of the estimates based on loo-CVC estimates is huge compared to those based on random projections. The upper right plot shows the corresponding distributions of the variance estimates from loo-CVC and random projections (based on $R = 100$ repetitions). The variances of the random projections follow a $\chi^2_n$ distribution (normalized by $n$) and hence have variance $2n^{-1}$. The variance of the loo-projections is far higher.

To explain this effect, the lower left plot shows the stability of the eigenvectors in the upper left plot, measured by the scalar product of sample eigenvectors and loo-sample eigenvectors (averaged over folds). One can see that there is a high instability of eigenvectors 8 and 9. The upper left plot shows that this is exactly where (i) the sample eigenvalues are very similar and (ii) the loo-CVC estimates deviate strongly from the population. To be precise, the loo-CVC corrections for the 8th (larger) and 9th (smaller) sample eigenvalues are underestimated and overestimated, respectively.

The plot to the lower right illustrates why this happens. It shows a subspace where both sample eigendirections $\tilde{b}_1$ and $\tilde{b}_2$ have very similar sample eigenvalues. Removing the yellow datapoint $x$, with
5.3. **Using Cross-Validation for Covariance Estimation**

![Figure 5.1. Illustration of the estimation error of loo-CVC.](image)
a high projection on \( \hat{\mathbf{v}}_2 \), does not considerably change the eigendirections. We have
\[
|\hat{\mathbf{v}}_1^T \mathbf{x}| < |\hat{\mathbf{v}}_2^T \mathbf{x}| \quad \Rightarrow \quad |\hat{\mathbf{v}}_{\text{loo}}^1 \mathbf{x}| < |\hat{\mathbf{v}}_{\text{loo}}^2 \mathbf{x}|.
\]
Removing the red data point \( \mathbf{x} \) with a high projection on \( \hat{\mathbf{v}}_1 \), on the other hand, changes the eigendirections completely: now \( \hat{\mathbf{v}}_{\text{loo}}^2 \) points in the direction of data point \( \mathbf{x} \) and we have
\[
|\hat{\mathbf{v}}_2^T \mathbf{x}| < |\hat{\mathbf{v}}_1^T \mathbf{x}| \quad \Rightarrow \quad |\hat{\mathbf{v}}_{\text{loo}}^1 \mathbf{x}| < |\hat{\mathbf{v}}_{\text{loo}}^2 \mathbf{x}|.
\]
It can be seen that no matter on which sample eigendirections the projection is higher, the projection on \( \hat{\mathbf{v}}_{\text{loo}}^2 \) is always higher than on \( \hat{\mathbf{v}}_{\text{loo}}^1 \). Therefore, the weaker direction has a strongly increased loo-CVC variance estimate.

A straightforward fix for this behavior is replacing loo by K-fold cross-validation. Using K-fold CV, it is highly unlikely that the entire set of hold-out data points has a high projection on the same eigendirection and is hence affected in the same way. This is shown in simulations after introducing isotonic regression.

**Reducing estimation error by isotonic regression** Some spectrum correction methods yield eigenvalues which are no longer ordered. This is not desired, as there is no reason to assume that a sample eigendirection with larger eigenvalue should have lower variance than one with smaller eigenvalue. Although sorting of the eigenvalue estimates is an option, a better alternative is the application of isotonic regression (Barlow et al., 1972). Isotonic regression searches for the decreasing sequence with minimum squared distance to the estimates:

\[
(\hat{\lambda}_1^\text{iso-loo-CVC}, \ldots, \hat{\lambda}_p^\text{iso-loo-CVC}) := \arg \min_{a \in \mathbb{R}^p} \sum_{i=1}^{p} (a_i - \hat{\lambda}_i^\text{loo-CVC})^2 .
\]

This quadratic program is reliably and quickly solved by freely available optimizers such as cvxopt, which was applied in the simulations and experiments below.

**Estimation error results** To compare CVC and NLS, the simulation shown in Figure 11 in (Ledoit et al., 2012) is repeated. There, 20%, 40% and 40% of the eigenvalues are set to 1, 3 and 10, respectively, the data is Gaussian and \( p/n = 1/3 \).

Unfortunately, there is no publicly available code for NLS and I was not able to exactly reproduce the NLS performance of Ledoit et al. (2012). My own implementation, using an open source implementation of Sequential Least Squares Programming (Perez et al., 2012; Kraft et al., 1988), is much slower and tends to get stuck in local optima. For precise estimation, it requires multiple initializations and obtained SEPRIALs are slightly worse for large \( p \).

Figure 5.2 shows that loo-CVC yields very bad results, which is in line with the argumentation of the last two paragraphs. Ledoit et al. (2012) reported much better results, which were only reproducible by incorporating a bug into the implementation of loo-CVC: the variance estimates of loo-CVC are calculated from projections \( \hat{\mathbf{V}}_{\text{loo}}^{\text{test}} \mathbf{x}_t \) (see eq. (5.10)). Removing the transpose, \( \hat{\mathbf{V}}_{\text{loo}}^{\text{test}} \mathbf{x}_t \), defines buggy-loo-CVC which exactly yields the result reported by Ledoit et al. (2012).

On the first glance, it is surprising that buggy-loo-CVC is better than the correctly implemented loo-CVC. This is explained by the fact that the simulation of Ledoit et al. (2012) is performed in the eigenbasis. In the eigenbasis, \( \mathbf{C} \) is diagonal, \( \hat{\mathbf{V}} \) is approximately the identity and hence \( \hat{\mathbf{V}}^T \approx \hat{\mathbf{V}} \). Using \( \hat{\mathbf{V}} \) instead of \( \hat{\mathbf{V}}^T \) has the advantage that the variance estimates do not suffer from the phenomenon illustrated in Figure 5.1. Randomly rotating the data before the analysis, buggy-loo-CVC breaks down completely and has much lower SEPRIAL than loo-CVC. All estimators besides buggy-loo-CVC are invariant w.r.t. the choice of basis.

The cross-validation with 10 folds in 10f-CVC yields much higher SEPRIALs than loo-CVC. Both CVC estimators profit from applying isotonic regression: iso-loo-CVC and iso-10f-CVC achieve competitive
5.3. USING CROSS-VALIDATION FOR COVARIANCE ESTIMATION

Figure 5.2. Estimation error of CVC estimators. Averaged over 50 repetitions.

Figure 5.3. Runtime comparison of NLS and CVC.

SEPRIALs compared to NLS. For a practical application, the small difference between NLS and iso-10f-CVC should be negligible.

5.3.3 Runtime comparison

In this section, the runtimes of CVC and NLS are compared. Both approaches are computationally more expensive than LW-Shrinkage:

- CVC requires multiple eigendecompositions, which are computationally expensive in high dimensions.
- NLS requires an expensive non-linear and non-convex optimization for the estimation of the population eigenvalues.

Ledoit et al. (2012) state that NLS is superior with respect to runtime, but they do not provide a numerical comparison.
CHAPTER 5. SPECTRUM CORRECTION

Table 5.1. CVC portfolio risk. Mean absolute deviations \(10^3\) (mean squared deviations \(10^6\)) of the resulting portfolios for the different covariance estimators and markets. * := significantly better than the other models at the 5% level, tested by a randomization test.

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>EU</th>
<th>HK</th>
</tr>
</thead>
<tbody>
<tr>
<td>aoc-Shrinkage</td>
<td>5.41 (67.0)</td>
<td>3.83 (36.3)</td>
<td>6.11 (71.8)</td>
</tr>
<tr>
<td>DVA-FA</td>
<td><strong>5.40</strong> (66.7)</td>
<td>3.84 (36.0)</td>
<td>6.12 (71.7)</td>
</tr>
<tr>
<td>CVC</td>
<td>5.40 (66.7)</td>
<td><strong>3.81</strong> * (35.7)</td>
<td><strong>6.10</strong> (71.8)</td>
</tr>
</tbody>
</table>

To compare the runtimes of NLS and CVC, the runtimes for both iso-loo-CVC and iso-10f-CVC is measured in the same simulation setting which led to Figure 12 in Ledoit et al. (2012). As I could not obtain the original program code by Ledoit and Wolf and my implementation is much slower, a proper comparison of runtimes on the same machine is not possible. The NLS results are taken from (Ledoit et al., 2012), the CVC results are calculated on a 2.3 GHz Intel i5 Macbook from 2011. The resulting runtimes should be roughly comparable.

Figure 5.3 shows that the runtime for NLS is higher than for the very slow iso-loo-CVC. The cost of iso-10f-CVC is negligible compared to the cost of NLS. Contrary to the hypothesis of Ledoit and Wolf, CVC is superior to NLS with respect to runtime.

5.4 Real world data

In this section, the iso-10f-CVC estimator is compared to factor models, aoc-Shrinkage and CV-Shrinkage in the experiments from Section 3.2.7. MI-BCI is not considered, as the CVC estimator is not designed for autocorrelated data.

Real world data I: portfolio optimization The portfolio analysis from Section 3.2.7 is performed for the CVC estimator. Table 5.1 compares the results to aoc-Shrinkage and DVA Factor Analysis. Although these approaches explicitly model the factor structure in the data, the CVC estimator is on par for the US and HK market and even significantly better for the EU market.

Real world data II: ISOLET and USPS The LDA and QDA analysis on the ISOLET and USPS data sets from Section 3.2.7 was repeated for the CVC estimator, PPCA, FA and DVA-FA. For the factor models, the number of factors is determined by the Bayesian Information Criterion. In addition, the optimal shrinkage intensity was optimized via cross-validation (CV-Shrinkage). To make cross-validation feasible for Quadratic Discriminant Analysis, a single shrinkage intensity is taken for all class covariance matrices.

The results in Table 5.2 show that the performance of the CVC estimator is comparable to aoc-Shrinkage. On these data sets, cross-validation outperforms the approaches which do not use label information. This shows clearly that the ESE is not a very good proxy for the classification accuracy. PPCA yields good results for QDA, where covariances are estimated on very little data. Besides, the factor models do not yield competitive performance. This indicates that the USPS and ISOLET data sets do not possess a very pronounced factor structure.

Real world data III: NU-ERP-BCI Table 5.3 shows results from the NU-ERP-BCI experiment. CVC outperforms aoc-Shrinkage, the best approach discussed in Section 3.2.7. As aoc-Shrinkage, CVC is not affected by the noise.
5.5. CHAPTER SUMMARY

Table 5.2. CVC, CV-Shrinkage and FM accuracies for classification tasks on ISOLET and USPS data.
* := significantly better than all other methods at the 5% level, tested by a randomization test.

<table>
<thead>
<tr>
<th></th>
<th>ISOLET 500</th>
<th>ISOLET 2000</th>
<th>ISOLET 5000</th>
<th>USPS 500</th>
<th>USPS 2000</th>
<th>USPS 5000</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVC-LDA</td>
<td>89.29%</td>
<td>93.46%</td>
<td>94.35%</td>
<td>83.49%</td>
<td>88.37%</td>
<td>89.76%</td>
</tr>
<tr>
<td>CV-Shrinkage-LDA</td>
<td>90.79%*</td>
<td>93.84%*</td>
<td>94.51%</td>
<td>86.75%</td>
<td>89.52%</td>
<td>90.24%</td>
</tr>
<tr>
<td>CVC-QDA</td>
<td>59.68%</td>
<td>83.34%</td>
<td>90.04%</td>
<td>81%</td>
<td>89.9%</td>
<td>91.1%</td>
</tr>
<tr>
<td>CV-Shrinkage-QDA</td>
<td>84.42%</td>
<td>92.06%</td>
<td>94.79%</td>
<td>91.4%*</td>
<td>95.47%*</td>
<td>96.14%*</td>
</tr>
<tr>
<td>aoc-Shrinkage-LDA</td>
<td>89.69%</td>
<td>93.42%</td>
<td>94.33%</td>
<td>83.95%</td>
<td>88.37%</td>
<td>89.77%</td>
</tr>
<tr>
<td>aoc-Shrinkage-QDA</td>
<td>59.51%</td>
<td>80.84%</td>
<td>87.35%</td>
<td>83.31%</td>
<td>89.4%</td>
<td>90.07%</td>
</tr>
<tr>
<td>PPCA-LDA</td>
<td>89.07%</td>
<td>91.55%</td>
<td>92.22%</td>
<td>82.3%</td>
<td>85.56%</td>
<td>86.32%</td>
</tr>
<tr>
<td>FA-LDA</td>
<td>88.79%</td>
<td>91.4%</td>
<td>91.9%</td>
<td>83.84%</td>
<td>86.65%</td>
<td>87.3%</td>
</tr>
<tr>
<td>DVA-FA-LDA</td>
<td>88.97%</td>
<td>91.45%</td>
<td>91.92%</td>
<td>84.24%</td>
<td>86.75%</td>
<td>87.33%</td>
</tr>
<tr>
<td>PPCA-QDA</td>
<td>83.93%</td>
<td>90.15%</td>
<td>91.58%</td>
<td>89.12%</td>
<td>78.7%</td>
<td>89.28%</td>
</tr>
<tr>
<td>FA-QDA</td>
<td>71.76%</td>
<td>84.99%</td>
<td>88.56%</td>
<td>76.47%</td>
<td>85.8%</td>
<td>87.24%</td>
</tr>
<tr>
<td>DVA-FA-QDA</td>
<td>56.25%</td>
<td>80.88%</td>
<td>88.37%</td>
<td>69.76%</td>
<td>83.27%</td>
<td>83.77%</td>
</tr>
</tbody>
</table>

Table 5.3. CVC, CV-Shrinkage and FM accuracies for classification tasks on the NU-ERP-BCI data.
Artificially injected noise in one electrode. * := significantly better than all other methods at the 5% level, tested by a randomization test.

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>10</th>
<th>30</th>
<th>100</th>
<th>300</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>CVC-LDA</td>
<td>93.74%</td>
<td>93.73%</td>
<td>93.73%</td>
<td>93.73%*</td>
<td>93.73%*</td>
<td>93.73%*</td>
</tr>
<tr>
<td>CV-Shrinkage-LDA</td>
<td>93.76%*</td>
<td>93.8%</td>
<td>93.74%</td>
<td>92.83%</td>
<td>89.92%</td>
<td>83.98%</td>
</tr>
<tr>
<td>aoc-Shrinkage-LDA</td>
<td>93.27%</td>
<td>93.27%</td>
<td>93.24%</td>
<td>92.88%</td>
<td>93.16%</td>
<td>93.19%</td>
</tr>
<tr>
<td>PPCA-LDA</td>
<td>91.13%</td>
<td>91.15%</td>
<td>91.11%</td>
<td>91.13%</td>
<td>91.07%</td>
<td>90.85%</td>
</tr>
<tr>
<td>FA-LDA</td>
<td>91.61%</td>
<td>91.6%</td>
<td>91.46%</td>
<td>91.14%</td>
<td>90.4%</td>
<td>90.35%</td>
</tr>
<tr>
<td>DVA-FA-LDA</td>
<td>91.96%</td>
<td>91.93%</td>
<td>91.17%</td>
<td>89.44%</td>
<td>90.05%</td>
<td>89.71%</td>
</tr>
</tbody>
</table>

CV-Shrinkage performs slightly better, yet the difference is not significant. For large noise levels, the optimal shrinkage intensity decreases and is no longer contained in the set of candidate values. This could easily be fixed, but it is shown here as a reminder that CV-Shrinkage depends on an appropriate set of candidate shrinkage intensities. As above, the factor models do not yield competitive performance: the NU-ERP-BCI data set does not possess a very pronounced factor structure either.

5.5 Chapter summary

This chapter presented covariance estimation methods which keep the sample eigenvectors and only modify the eigenvalues. The state-of-the-art is Nonlinear Shrinkage (NLS), a highly complex method from Random Matrix Theory which minimizes the expected squared error. I proposed a cross-validation-based covariance (CVC) estimator based on isotonic regression and 10-fold cross-validation which optimizes the same loss function and yields competitive results at greatly reduced complexity and runtime.

Performance On all data sets considered, the CVC estimator yields better or equal performance in comparison to state-of-the-art covariance estimators based on factor models or shrinkage and simulations show that the estimation errors of CVC and NLS are on the same level. Only CV-Shrinkage, which takes the label information into account, yields higher accuracies in real world applications of Linear or Quadratic Discriminant Analysis.
Runtime  Both NLS and CVC are computationally more complex than LW-Shrinkage. The non-linear and non-convex optimization required by NLS is very time-consuming. A comparison in the simulation setting considered by Ledoit & Wolf (2012) showed that the cost of iso-loo-CVC is significantly smaller, although it requires \( n + 1 \) eigendecompositions. Using 10 folds greatly reduces the computational cost and renders CVC applicable in practice.
Chapter 6

Summary and Outlook

The sample covariance matrix is the standard estimator of the population covariance: it is simple, fast and converges to the population covariance in the limit of infinitely many observations. The drawbacks of the sample covariance become apparent if the dimensionality is of the same order as the number of observations: then, the eigenvalues of the sample covariance matrix are highly biased, the condition number is large and the inversion of the matrix is numerically unstable.

Motivated by the large number of algorithms which require accurate covariance matrix estimates, high-dimensional covariance matrix estimation has become a wide research field with many subfields and a vast literature. In this thesis, I have discussed the history as well as the state-of-the-art of the field. Furthermore, I have presented my contributions which yield improvements upon the state-of-the-art in all three subfields of structured estimation, regularized estimation and spectrum correction methods.

Structured estimation Chapter 2 focused on models with low intrinsic dimensionality. It turned out that the widely used Factor Analysis model suffers from a bias similar to that of the sample covariance matrix: the variance in factor directions and low variance directions orthogonal to the factors is over- and underestimated, respectively. I proposed Directional Variance Adjustment (DVA) Factor Analysis, an algorithm which reduces the bias and yields improved estimates of the covariance matrix. In a detailed portfolio simulation based on real world stock market data from three different stock exchanges, DVA Factor Analysis yielded significantly better performance than standard Factor Analysis, state-of-the-art regularized estimators and other competing approaches from quantitative finance.

Regularized estimation Chapters 3 and 4 discussed shrinkage, the most popular regularized covariance estimation technique. In a first step (Section 3.2), I clarified the implicit assumptions on the population covariance structure of Ledoit-Wolf-Shrinkage: the presence of strong eigendirections, which I showed to be a highly relevant case, is ruled out. After generalizing the proof of consistency to arbitrary covariance structures and showing that strong eigendirections cause LW-Shrinkage to converge to the sample covariance matrix, I proposed automatic orthogonal complement Shrinkage, which takes the strong eigendirections into account and yielded significant improvements in classification tasks on handwritten digits, spoken letter recognition and MSM-ERP-BCI data as well as portfolio optimization.

An important aspect of LW-Shrinkage is its vulnerability to autocorrelation of the observations. Section 3.3 introduced the Sancetta-Estimator, an extension of LW-Shrinkage to autocorrelated data. I showed that the Sancetta-Estimator is biased and proposed a theoretically and empirically superior shrinkage estimator which is consistent under autocorrelation and has strongly reduced bias and lag parameter sensitivity.

Chapter 4 discussed the extension of shrinkage to multiple shrinkage targets, which allows for many interesting applications which go beyond regularization, including transfer learning. I proposed Multi-Target Shrinkage (MTS) in a general framework, independently of a specific estimator, and proved its
consistency. For shrinkage of sample mean and sample covariance, the chapter provided a detailed theoretical analysis and extensive simulations. In addition, it presented two real world applications in which MTS is superior to standard LW-Shrinkage.

Spectrum correction The topic of Chapter 5 was improved covariance estimation by replacing the sample eigenvalues with improved estimates. I contributed to spectrum correction by proposing a cross-validation based covariance (CVC) estimator which yields a performance which is competitive to the state-of-the-art technique Nonlinear Shrinkage (NLS). While NLS has excellent theoretical properties in the large-dimensional limit, the algorithmic instability of the eigendecomposition makes the theoretical analysis of CVC hard. The simulation study showed clear advantages of CVC over NLS with respect to numerical stability and computational cost. An empirical comparison with the structured and regularized estimators discussed in the Chapters 2 and 3 on stock market data, handwritten digits, spoken letters and BCI data gave evidence that CVC is the best general purpose covariance estimator available.

Estimation error vs. performance An important aspect of covariance estimation is the difference between estimation error and performance. While covariance estimators minimize measures of estimation error, the quantity of interest for a practitioner is the performance in an application, e.g. classification accuracies or portfolio risk. The analysis of real world data showed that in some cases improvements of estimation error even reduced the performance (see, e.g., Table 3.4). For the considered data sets, the direct optimization of performance by cross-validation of the regularization strength yielded higher classification accuracies than the best covariance estimators, albeit at a high computational cost.

For Linear Discriminant Analysis as a model application, Appendix A analyzed the interdependency between the generative model of the data and different covariance estimators. There, it was shown that LW-Shrinkage only yields high classification accuracies for a specific localization of the discriminative information, while CV-Shrinkage can use label information to adapt to different localizations and CVC is flexible enough to yield good estimates in low and high variance regions at the same time. Below, Section 6.2 concludes with a discussion of the advantages and disadvantages of all covariance estimators presented in the thesis and gives situation-specific recommendations.

6.1 Open questions and future work

The field of covariance estimation is a rather mature field. A large set of covariance estimators exists and is used in practice. Yet, in the light of the research presented in this thesis, a set of open questions remains.

Structured estimation In this thesis, only the assumption of low intrinsic dimensionality was considered. Compared to Factor Analysis and DVA Factor Analysis, CVC yields as good or superior performance, even if the data exhibits a clear factor structure. Settings in which other structural assumptions make sense are also of high interest. The work in this thesis directly motivates two questions:

- **Do specialized approaches have an advantage?** In portfolio optimization, the general purpose covariance estimator CVC performs on par with estimators which explicitly model the factor structure. It would be interesting to analyze whether CVC, in the corresponding settings, is also on par with estimators which make assumptions about sparsity, sparsity of the inverse or band-structure.

- **Are there transferable ideas?** The estimators presented in the chapters on regularization and spectrum correction do not make any structural assumptions. It is possible to

  (i) modify some of the regularized approaches/spectrum correction methods by imposing constraints and thereby developing new structured estimators.
(ii) transfer ideas used in the design of the regularized estimators/spectrum correction methods to existing structured estimation approaches.

**Multi-Target Shrinkage** As shown in Section 4, the extension of shrinkage to multiple targets allows for interesting applications. There is a lot of potential for further applications, which have to be developed in cooperation with practitioners. An aspect of MTS which should be investigated further is the whitening trick:

- **Incorporating label information**: performance gains could be obtained (i) by incorporating label information into the weighting of the different dimensions and (ii) by adaptively whitening only to an extent which sufficiently reduces the variance of the shrinkage estimates.
- **Theoretical analysis**: the whitening trick has been introduced as a heuristic. A theoretical analysis could clarify the conditions under which whitening reduces the variance of the MTS shrinkage intensity estimates.

**Spectrum correction** In the last decade, shrinkage has become the most widely used covariance estimation technique. Due to obvious shortcomings, spectrum correction techniques as proposed by Stein and Dey have not been adopted by practitioners. Promising alternatives are NLS and CVC, they are more flexible and more powerful, and simulation results show that the increased flexibility comes at a moderate cost w.r.t. increased variance. Three aspects require further research:

- **High computational complexity**: a disadvantage of CVC and NLS is the relatively high computation time, which is close to (CVC) or above (NLS) the level of CV-Shrinkage. Further research could reduce optimization time for NLS or yield a more efficient update strategy for CVC.
- **Autocorrelation**: NLS and CVC both assume i.i.d. data. For NLS, an extension to autocorrelated data could be based on estimating the effective sample size. For CVC, training and test folds have to be independent. As written above, for an autocorrelation decay which is fast relative to the sizes of the folds, folds based on consecutive data points are approximately independent.
- **Sample eigenbasis**: all spectrum correction approaches rely on the sample eigenbasis. This gives room for improvement, e.g. by taking a robust estimator for the eigenbasis.

**Relation between estimators and applications** There is still little knowledge about the interdependency between covariance estimators and their applications. In my opinion, this aspect should be investigated in detail:

- **Simulation studies**: the study in Appendix A is, to my knowledge, the first systematic analysis on the interdependencies between different covariance estimators and an application, in this case Linear Discriminant Analysis. A similar analysis for other applications could yield a better understanding of the limitations of each estimator and strengthen the empirical basis of the recommendations in the next section.
- **Real world data studies**: a large comparison of all relevant estimators on all relevant data sets, similar to the large comparison of classifiers by Fernández-Delgado et al. (2014), would be of interest.

### 6.2 Which estimator to use in practice

For practitioners, there is bad news: it is not possible to say that one of the covariance matrix estimators is simply better than all the others. This is related to the **no free lunch theorem** (Wolpert, 1996; Wolpert & Macready, 1997). The estimators discussed in this thesis are summarized in Table 6.1. Their individual differences lie in their computational cost and performance, and to make things even more complicated, these aspects in turn depend on the specific domain of interest.
Table 6.1. Covariance matrix estimators.

<table>
<thead>
<tr>
<th>Estimator</th>
<th>Comp. cost</th>
<th>Properties</th>
</tr>
</thead>
<tbody>
<tr>
<td>sample covariance</td>
<td>very low</td>
<td>+ simple and fast</td>
</tr>
<tr>
<td></td>
<td></td>
<td>consistent for $n \to \infty$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>− high systematic error in the spectrum</td>
</tr>
<tr>
<td></td>
<td></td>
<td>high condition number for $p \approx n$</td>
</tr>
<tr>
<td></td>
<td></td>
<td>rank-deficient for $p &lt; n$</td>
</tr>
<tr>
<td><strong>Structured estimators</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Probabilistic PCA</td>
<td>low/medium</td>
<td>+ few parameters</td>
</tr>
<tr>
<td></td>
<td></td>
<td>closed-form factors</td>
</tr>
<tr>
<td></td>
<td></td>
<td>− isotropic noise often too restrictive</td>
</tr>
<tr>
<td>Factor Analysis (FA)</td>
<td>medium/high</td>
<td>+ higher flexibility than PCA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>− non-convex problem</td>
</tr>
<tr>
<td></td>
<td></td>
<td>slow iterative algorithm</td>
</tr>
<tr>
<td>DVA Factor Analysis</td>
<td>very high</td>
<td>+ better performance than FA</td>
</tr>
<tr>
<td></td>
<td></td>
<td>− slower than FA</td>
</tr>
<tr>
<td><strong>Shrinkage</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LW-Shrinkage</td>
<td>low</td>
<td>+ very fast</td>
</tr>
<tr>
<td></td>
<td></td>
<td>good theoretical properties</td>
</tr>
<tr>
<td></td>
<td></td>
<td>autocorrelation-consistent extension possible</td>
</tr>
<tr>
<td></td>
<td></td>
<td>− no performance guarantees</td>
</tr>
<tr>
<td></td>
<td></td>
<td>structural assumption</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dominated by high-variance directions</td>
</tr>
<tr>
<td>aoc-Shrinkage</td>
<td>medium</td>
<td>+ lower ESE than LW-Shrinkage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>moderate increase in computational cost</td>
</tr>
<tr>
<td></td>
<td></td>
<td>− slower than LW-Shrinkage</td>
</tr>
<tr>
<td>CV-Shrinkage</td>
<td>high</td>
<td>+ direct optimization of performance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>excellent performance in practice</td>
</tr>
<tr>
<td></td>
<td></td>
<td>− slow</td>
</tr>
<tr>
<td></td>
<td></td>
<td>requires label information</td>
</tr>
<tr>
<td><strong>Spectrum correction</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonlinear Shrinkage (NLS)</td>
<td>very high</td>
<td>+ good theoretical foundation</td>
</tr>
<tr>
<td></td>
<td></td>
<td>low estimation error</td>
</tr>
<tr>
<td></td>
<td></td>
<td>higher flexibility than LW-Shrinkage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>− involved &amp; unstable numerical optimization</td>
</tr>
<tr>
<td></td>
<td></td>
<td>very slow</td>
</tr>
<tr>
<td>CVC</td>
<td>medium/high</td>
<td>+ good performance</td>
</tr>
<tr>
<td></td>
<td></td>
<td>easy to implement</td>
</tr>
<tr>
<td></td>
<td></td>
<td>faster than NLS</td>
</tr>
<tr>
<td></td>
<td></td>
<td>higher flexibility than LW-Shrinkage</td>
</tr>
<tr>
<td></td>
<td></td>
<td>− slow</td>
</tr>
</tbody>
</table>

**Computational cost**  If estimation speed is a major concern, as for example in online applications, approaches based on cross-validation, involved numerical optimization and resampling are infeasible.

Although the sample covariance is clearly the least time consuming method, LW-Shrinkage is nearly on par. The time demand of PPCA and aoc-Shrinkage can get prohibitive when confronted with high-dimensional settings, where the eigendecomposition of the sample covariance is very expensive. When the dimensionality is not extremely high or the eigendecomposition is needed in other stages of the analysis pipeline, aoc-Shrinkage can yield better performance at the cost of a moderate increase in computation.
The CVC-estimator is slower given that it requires multiple eigendecompositions, yet its good performance makes it an interesting alternative to cross-validation. Nonlinear Shrinkage is much slower and yields approximately the same result as CVC.

Factor Analysis and DVA are based on iterative algorithms which are slow. For time-critical applications, especially DVA is much too slow.

**Performance** The standard measures of estimation error like ESE are not relevant in practice. Practitioners are interested in application-dependent performance measures, e.g. the realized portfolio risk or the accuracy of a classification algorithm.

Without knowledge of the application domain, it is impossible to know a priori which estimator will yield the best performance. Only simulations on empirical data yield evidence for hypotheses such as LW-Shrinkage yields good accuracies in ERP-BCI. An exception is CV-Shrinkage, which does this evaluation directly and is hence able to perform well on most data sets.

An empirical validation allows to select the best estimator from a set of candidates, but this is rather not satisfying from a research perspective. Analyzing the domain-dependency is complex, as it consists of two interdependent aspects:

1. The data generating process results in a covariance structure which has a strong influence on covariance estimation.
2. The applied machine learning method or statistical procedure depends in a very specific way on the estimation errors.

Only knowledge of both aspects and knowledge of how these aspects relate to the properties of the covariance estimators allow for an informed decision for a specific covariance estimator.

For Linear Discriminant Analysis, it was shown in Appendix A what such an analysis could look like. In LDA, it is decisive (i) whether the discriminative information lies in the low- or high-variance regions of the spectrum and (ii) how dispersed the eigenvalues are. Each covariance estimator changes the sensitivity of LDA with respect to these parameters.

**Guidelines for the choice of a covariance estimator** In the last paragraphs, it was explained why there is no simple recommendation for the best covariance estimator. Yet there are some guidelines. Some of these may appear trivial to the informed reader, but are included for the sake of completeness.

1. Domain knowledge has to be acquired in order to make an informed decision. Otherwise it is not possible to know whether the optimized measure of estimation error is a good proxy for the performance in the application under consideration.
2. In most cases, the sample covariance is a bad choice. Only if data is abundant or the application is insensitive to systematic errors in the spectrum, it yields competitive results.
3. LW-Shrinkage is nearly as fast as the sample covariance and often yields good results in practice. We have to be aware that there is no guarantee for a good performance. A good choice of shrinkage target can significantly improve the performance.
4. In general, aoc-Shrinkage tends to improve upon LW-Shrinkage. If the increase in computational cost is not a problem, it should be used.
   *Note:* As the improvement in comparison to LW-Shrinkage is an improvement in estimation error, there is no guarantee for an improved performance in an application.

5. CVC tends to outperform LW- and aoc-Shrinkage, albeit at a higher computational cost. If computation time is not an issue, CVC should be used instead of LW- or aoc-Shrinkage.
6. NLS optimizes the same objective as CVC and yields similar performance. With NLS, it is also possible to optimize the precision matrix, but the systematic analysis in Appendix A indicates that this does not yield improved performances in practice. The NLS optimization problem is highly complex and no publicly available implementation exists, while CVC can be implemented in a few lines of code. CVC is clearly superior with respect to computational costs.

7. If there is enough data for reliable cross-validation, if labels are available and if computation time is not an issue, CV-Shrinkage is superior to LW and aoc-Shrinkage as it directly optimizes the quantity of interest. The systematic analysis has shown that in standard settings it also outperforms CVC. CVC is superior to CV-Shrinkage when the labels contain little information or when the relevant information is spread over directions of high and low variance. In the real world applications considered in this thesis, CVC did not outperform CV-Shrinkage.

8. If the data has clear factor structure, factor models tend to perform well.
   (a) If noise is isotropic or there is very little data, probabilistic PCA is a good choice.
   (b) for anisotropic noise (and sufficient data to estimate this asphericity), Factor Analysis is superior.
   (c) DVA Factor Analysis improves upon Factor Analysis, albeit at a higher computational cost.

   Probabilistic PCA and Factor Analysis are based on maximum likelihood. As for LW-Shrinkage, there is no guarantee for a good performance in practice. For all considered real world applications, CVC yields better or equal performance at comparable/lower computational cost than standard/DVA Factor Analysis, even if there is a clear factor structure in the data.

9. If the data set is autocorrelated, this has to be taken into account. Autocorrelation violates the assumptions of LW-Shrinkage, aoc-Shrinkage, CVC and NLS. Instead, the bias-corrected autocorrelation consistent LW-Shrinkage approach proposed in Section 3.3 should be used.

   If the decay of the autocorrelation is fast relative to \( n/10 \), iso-10f-CVC should still yield good results, as the folds are approximately independent.

10. If multiple target candidates and/or additional data sets from a similar distribution exist, Multi-Target Shrinkage yields superior performance compared to LW-Shrinkage.

   Note: As the improvement over LW-Shrinkage is an improvement in estimation error, there is no guarantee for an improved performance in an application.
Appendix A

Estimation Error vs. Performance

In this thesis, it is often stated that measures of estimation error like ESE, likelihood, Stein’s loss or PRIAL are only proxies for the application-dependent quantities of true interest: practitioners in finance, for example, are interested in the variance of portfolios and researchers in the field of Brain-Computer Interfaces aim at high classification accuracies.

Figure A.1 illustrates the difference between the optimization of estimation error in LW-Shrinkage/iso-10f-CVC and the optimization of performance in CV-Shrinkage. The figure shows the application of Linear Discriminant Analysis to a 3-class problem where weak regularization is optimal. A non-discriminative noise dimension is added, the x-axis shows the noise magnitude.

It can be seen that the accuracies (upper plot) of standard LDA, CV-Shrinkage-LDA and CVC-LDA are independent of the noise. LW-Shrinkage-LDA, on the other hand, shows a weird behavior and is inferior to CV-Shrinkage-LDA.

The picture is very different for the ESEs of the covariance estimators: the PRIAL (upper middle) of LW-Shrinkage is always positive while CV-Shrinkage incurs a squared error up to ten times as high as the sample covariance.

The shrinkage intensities (lower middle) show that for increasing noise, $\lambda^{\text{LW}}$ converges to 0.01, while $\lambda^{\text{CV}}$ is larger for small noise but continues to decay when $\lambda^{\text{LW}}$ has converged. Yet, to understand the accuracies, one has to consider the regularization $\lambda \text{Tr} (\tilde{S})/p$ (lower plot): with increasing noise, there is an increase in the values on the diagonal of the target, $\text{Tr} (\tilde{S})/p$. This can compensate for a decreasing shrinkage intensity.

- In LW-Shrinkage, the regularization first decreases, which causes a performance drop. It then increases, leading to good performance. Finally, regularization becomes much too strong, which again results in suboptimal classification accuracies.

- CV-Shrinkage on the other hand, decreases the shrinkage intensity such that the obtained regularization remains approximately constant.

In this appendix, the relationship between LDA and covariance estimation is analyzed in detail. LDA has the advantage of being mainly governed by two factors: the shape of the covariance and the orientation of the discriminative information relative to the covariance. In principle, similar analyses could be performed for all applications of covariance matrices (see Section 1.3), however, this is beyond the scope of this dissertation.

The simulation results in the following sections will show that one has to be cautious whenever a covariance estimator based on a measure of estimation error is applied. Without an empirical validation,

---

1The experiment is based on experiment 4 in Friedman (1989), with the dimensionality increased to $p = 100$.
2The setting in the simulation differs from the theoretical analysis of LW-Shrinkage and covariance structure in Section 3.2: there, $\text{Tr} (\tilde{S})/p$ is bounded and $\lambda \to 0$ implies that there is no regularization in the limit.
3For noise injection, ooc-Shrinkage also works perfectly. See the NU-ERP-BCI experiment in Section 3.2.7.
Figure A.1. Shrinkage failure in the presence of a noise channel.
A.1. DISCRIMINATIVE INFORMATION AND COVARIANCE

Figure A.2. 2D illustration: average predictions. The color coding indicates the probability that an LDA classifier based on five observations per class classifies a point as red or blue. Dark blue and dark red indicate 100% probability of being classified as class blue and class red, respectively.

there is no guarantee that an estimator with low estimation error also yields high performance. This appendix is based on an article in preparation (Bartz, 2015b).

Appendix A outline In Section A.1, an intuitive explanation for the behavior of LDA under regularization is given. Section A.2 contains a description of the systematic analysis of LDA. The results of the analysis are presented in Section A.3.

A.1 Discriminative information and covariance

For Linear Discriminant Analysis, the orientation of the covariance relative to the location of the sample means is decisive.

Figure A.2 illustrates this in a two-dimensional setting with five observations for each of the two classes red and blue. The upper and lower row show two different settings: for the upper row, the mean differences lie in the direction of high variance, as illustrated by the constant likelihood contours of the distributions of the two classes. For the lower row, the mean differences lie in the direction of low variance.

The two columns show two different classification approaches. The left column shows standard LDA, equivalent to Shrinkage-LDA with $\lambda = 0$, while the right shows the Nearest Centroid Classifier, equivalent to Shrinkage-LDA with $\lambda = 1$.

The color coding shows classification probabilities for the two classes. Interestingly, the picture is very different for the two settings: if the high-variance direction is discriminative, the Nearest Centroid Classifier is much better, if the low-variance direction is discriminative, LDA gives better decision boundaries.

To understand this, consider Figure A.3. If the direction of high variance is discriminative, relatively small errors in shape of the covariance estimate can lead to highly suboptimal decision boundaries. If the direction of low variance is discriminative, errors in the mean in the direction of high variance adversely affect nearest centroids.

The illustrations deliver an important message: from the point of view of covariance estimation, the two
settings are indistinguishable. Although very different covariance estimators are optimal, the covariance estimates of methods like LW-Shrinkage, NLS and CVC will not differ. This is especially dramatic for LW-Shrinkage-LDA which yields an interpolation between standard LDA and nearest centroids and cannot be optimal in both settings.

A way out is given by CV-Shrinkage. CV-Shrinkage takes the label information into account and is able to choose a small or large shrinkage intensity when low- or high-variance directions are discriminative, respectively.

**Discriminative information and covariance in more than two dimensions** In the simulation above, either the low- or the high-variance direction was discriminative. Below, in the simulations for higher dimensionality, mainly three settings are of interest:

- **high-variance directions are discriminative**: it is often assumed that the directions of high variance are informative. This does not mean that the Nearest Centroid Classifier is optimal, because the discriminative high-variance directions may vary in strength. An alternative definition of this setting is that the discriminative directions are those which dominate the Frobenius norm of the covariance matrix.

- **low-variance directions are discriminative**: low-variance directions are all directions which are dominated by the strongest eigendirections. Note that this definition does not only include the directions of lowest variance, which are often discarded in practice as they do not carry discriminative information. An alternative definition of this setting is that the directions which dominate the Frobenius norm of the covariance matrix are not discriminative.

- **spread-out discriminative information**: in this setting, the discriminative information is not concentrated in the subspace of high or low variance.

*Braun et al. (2008)* studied the relationship between discriminative information and directions of high variance for the case of kernel principal component analysis. They proved that the discriminative information is contained in the leading components if the kernel matches the learning problem. Note that
A.2. SIMULATION DESIGN

Figure A.4. SNR and eigenvalues in the main simulation and variation 1.

this theoretical result does not pose a restriction on the location of the discriminative information in
the LDA setting: the linear kernel, tightly related to the covariance matrix, only matches the learning
problem if the discriminative information is contained in the leading components. If the discriminative
information is not restricted to the leading components, the theoretical result is not applicable without
further assumptions.

A.2 Simulation design

The last section has shown that the orientation of the covariance relative to the discriminative directions
plays an important role in Linear Discriminant Analysis. This section contains a systematic analysis.

Main simulation For the main simulation, Gaussian data with two classes is generated \((p = 100, n = 200)\). The eigenvalues \(\gamma_i\) are chosen logarithmically spaced between \(10^{-\alpha}\) and \(10^\alpha\) and values of \(\alpha\) between 0 and 2 are considered. For \(\alpha = 0\) the spectrum is flat, for \(\alpha = 2\) the spectrum is heavily tilted (see Figure A.4).

The mean of class A is set to \(\mu^A = 0\), while the entries in \(\mu^B\) are defined in the following way: \(\mu^B_i/c_{90}\%/\sqrt{\gamma_i}\) are logarithmically spaced between \(10^\beta\) and \(10^{-\beta}\), where \(c_{90}\%\) is chosen such that the Bayes optimal classifier obtains 90\% accuracy.

For \(\beta\), values between -1 and 1 are considered. For \(\beta < 0\) the Signal-to-Noise ratio is high in the
directions of low variance, for \(\beta = 0\) the SNR is identical in all directions and for \(\beta > 0\) the SNR is high
in the directions of low variance (see Figure A.4).

Variation 1: low Bayes-optimal classification rate As a variation of the main simulation, the
Bayes-optimal classification rate of was set to 60\%. This increases the noise in the cross-validation of the
optimal shrinkage intensity. Apart from that, the simulation was identical to the previous one.

Variation 2: non-monotonous SNR The second variation considered a slightly artificial case: the
information is concentrated either in the medium-variance range or only in the high and low-variance
directions. Variation 2 only differs in the definition of \(\mu^B\). The first and the second half of the entries
in $\mu^B$ are defined such that the corresponding $\mu^B_i / c_{99%} / \sqrt{n}$ are logarithmically spaced between $10^\beta$ and $10^{-\beta}$ and $10^{-\beta}$ and $10^\beta$, respectively. Figure A.5 shows the distribution of the eigenvalues and the SNR.

A.3 Simulation results

Figure A.6 shows the results of the main simulation. The systematic analysis yields answers to a set of questions:

- **How good are sample and population covariance?**
  Figure A.6 (first row) shows that in this setting, the population covariance achieves an accuracy of approximately 87%, independently of $\alpha$ and $\beta$. Compared to the Bayes optimal classifier, 3% are lost due to the estimation error in the sample means. The sample covariance is about 8% worse than the population covariance.

- **Is the precision oracle better than the (spectrum) oracle?**
  In the last chapter, two optimal estimators in the sample eigenbasis were discussed: the (spectrum) oracle eq. (5.7) has minimum squared error with respect to covariance, the precision oracle eq. (5.12) has minimum squared error with respect to the precision matrix. Figure A.6 (second row) shows that the results are very similar. Classification accuracies differ by less than 1%. If high-variance directions are discriminative or the information is spread evenly, the oracle is superior. If directions of low variance are discriminative, the precision oracle is superior.

To understand this behavior, consider the oracles in the population eigenbasis. Then, we have from eq. (5.7) and eq. (5.12):

$$\gamma^*_i = \sum_j \hat{\gamma}^2_j \gamma_j \quad \text{and} \quad \gamma^*_i = \frac{1}{\sum_j \hat{\gamma}^2_j \gamma_j}.$$  \hfill (A.1)

The oracle is a weighted arithmetic mean of the eigenvalues, while the precision oracle is a weighted harmonic mean. The harmonic mean is always smaller than the arithmetic mean, but the ratio differs for large and small eigenvalues. Figure A.7 shows that the precision oracle yields, relative to
Figure A.6. Main simulation: systematic comparison of classification accuracies.
the oracle, very small eigenvalues for the small sample eigenvalues. Hence it favors the directions of small variance and is superior when these directions are discriminative.

- **Does the CVC estimator achieve performance comparable to the oracle?**
  Figure A.6 (middle row) shows that the differences in accuracy are smaller than 1%. If directions of high variance are discriminative, the difference is close to zero, it is largest for discriminative low-variance directions.

- **How does LW-Shrinkage compare to CVC?**
  Figure A.6 (second last row) shows that both estimators are on par when the spectrum is flat. For a strongly tilted spectrum, LW-Shrinkage performs slightly better when the directions of high variance are discriminative. When the directions of low variance are discriminative, CVC performs much better, in the extreme case the difference in accuracies is at 20%.

- **How does CV-Shrinkage compare to CVC?**
  Figure A.6 (last row) shows that CV-Shrinkage is very good for LDA. By using label information,
CV-Shrinkage is able to tune directly to the part of the spectrum where the discriminative information lies. Thereby it yields a performance increase of up to 3% relative to CVC. CVC is better if the discriminative information is spread evenly over all parts of the spectrum. There are settings less favorable for CV-Shrinkage:

- **lower accuracy of the Bayes-optimal classifier**: when the Bayes-optimal classifier only obtains an accuracy of 60% in variation 1, the label information used by cross-validation is less informative. Figure A.8 shows that in this case, CV-Shrinkage less accurately finds the optimal shrinkage intensity: the parameter range in which CVC outperforms CV-Shrinkage gets larger.
- **non-monotonous SNR**: Figure A.9 shows the comparison between CV-Shrinkage and CVC for variation 2, where the SNR is non-monotonous. One can see that in this case the shrinkage framework is not flexible enough. This setting is not very realistic but instructive.

**Limitations of the simulations** Although the LDA setting is relatively simple, the analysis presented in this section cannot be complete: it focused on the eigenvalue dispersion and the orientation of the covariance and had to neglect other aspects:

- **Concentration $c$, dimensionality $p$ and number of observations $n$**: in the above analysis, $p = 100$ and $n = 200$ were fixed. At lower concentration $c = n/p$ higher regularization is optimal. At smaller $n$ all estimators get less precise.

- **Shape of the spectrum**: the logarithmic spacing of the population eigenvalues, displayed in Figure A.4 and A.5, is just one possible parametrization. Besides other functional forms there could also be a factor structure with single eigenvalues clearly separated from the bulk.

- **Shape of the SNR**: the locations of the means which were chosen for the simulations yield the SNRs displayed in Figure A.4 and A.5. Other locations of the means leading to differently shaped SNR curves are possible.

- **Effect of the distribution**: real world data is often non-Gaussian. It would be interesting to analyze the effect a distribution with heavy tails has on the results.

- **Number of classes**: the systematic analysis only considered two classes. An extension to multiple classes would greatly increase the degrees of freedom in the analysis.
Relation to experiments on real world data  In Sections 3.2.7 and 5.4, Linear Discriminant Analysis has been applied to the MSM-ERP-BCI, ISOLET and USPS data sets. The systematic analysis gives us insights into the results reported there.

In general, the discriminative information in these experiments seems to be restricted to the high-variance dimensions: LW-Shrinkage yields good results, although CVC tends to be a bit better. In addition, the Bayes-optimal classification rate is relatively high in all experiments. These two facts explain why the best results are obtained by CV-Shrinkage.

In none of the experiments the discriminative information is spread evenly over large regions of the spectrum, the setting in which CVC would outperform CV-Shrinkage. This does not rule out the existence of such settings in other application domains.

A.4 Appendix summary

The performance in a practical application is the main uncertainty in covariance estimation research. All covariance estimators considered in this thesis optimize a measure of estimation error, which may or may not be a good proxy for the application-dependent performance measure.

In this appendix, light was shed on a specific application of covariance matrices. Linear Discriminant Analysis is a widely used classification technique, which is often enhanced by LW-Shrinkage. It was shown that in LDA the locations of the sample means, relative to the orientations of the covariance, are decisive. They have a huge influence, for example, on the optimal shrinkage intensity w.r.t. to classification accuracy. Optimizing a proxy like ESE does not incorporate information on the location of the class means: the covariance estimates of LW-Shrinkage, NLS and CVC are location-invariant.

LW-Shrinkage is only optimal if the discriminative information is in the high-variance region of the covariance. The squared error objective, in combination with the limited flexibility of shrinkage, leads to large relative errors in the low-variance part of the spectrum. CVC, on the other hand, is more flexible and achieves a good fit for both small and large eigenvalues.

CV-Shrinkage takes information on the locations into account and can hence optimize for the part of the spectrum in which the discriminative information lies. As long as the information is restricted to a specific region of the spectrum, CV-Shrinkage outperforms all covariance estimators. Only if the information is not restricted to a specific region, CV-Shrinkage is not flexible enough and is outperformed by spectrum correction.
Appendix B

Proofs and Derivations for Shrinkage and Structure

B.1 Proof of Theorem 3.1 (Largest eigenvalue growth rate)

Proof. Under assumption \((LW2)\), we have

\[
K_2^{(LW2)} \geq \frac{1}{p_n} \sum_{i=1}^{p_n} \mathbb{E} \left[ (y_n^i)^8 \right] \geq \frac{1}{p_n} \sum_{i=1}^{p_n} \mathbb{E}^4 \left[ (y_n^i)^2 \right] \geq \frac{\gamma_1^4}{p_n},
\]

\[
\iff \gamma_1^4 \leq p_n^{3/4} K_2^{1/4},
\]

\[
\implies \gamma_1 = O \left( p_n^{1/4} \right),
\]

which is the rate from the Theorem.

To show that this is the lowest bound which holds for all sequences, let us assume (i) \(p_n = n\) and (ii) \(y_n^i\) is a zero mean Gaussian variable with \(\mathbb{E}[(y_n^i)^2] = \sqrt{p_n}\), while \(y_n^{p_n+1} = 0\). For this sequence, \((LW2)\) holds:

\[
\frac{1}{p_n} \sum_i \mathbb{E}[(y_n^i)^8] = 105.
\]

For the largest eigenvalue we have \(\gamma_1^n = \sqrt{p_n}\), which attains the above rate and thereby proves that no lower rate exists for all sequences.

B.2 Proof of Theorem 3.2 (Dispersion growth rate)

Proof. Under Assumption \((LW2)\), we have

\[
d_n := \frac{1}{p_n} \sum_i (\gamma_i - \sum_j \gamma_j)^2 \leq \frac{1}{p_n} \sum_i \gamma_i^2 \leq \sqrt{\frac{1}{p_n} \sum_i \gamma_i^4} \leq \sqrt{\frac{1}{p_n} \sum_i \mathbb{E}[(y_n^i)^8]}^{(LW2)} \leq \sqrt{K_2},
\]

\[
\iff d_n = O(1),
\]

which is the rate from the Theorem.

To show that this is the smallest rate which bounds all sequences, let us consider a sequence of Gaussian data with

\[
\gamma_i = \begin{cases} 
\zeta & \text{for } i \leq p_n/2 \\
0 & \text{for } i > p_n/2.
\end{cases}
\]
There is a $\zeta$ for which (LW2) is fulfilled:

$$\frac{1}{p_n} \sum_i \mathbb{E}[(y_i^*)^8] = \frac{1}{p_n} 105 \sum_i \gamma_i = \frac{105}{2} \zeta \leq K_2.$$ 

For the dispersion, we have

$$d_n = \frac{1}{p_n} \sum_i (\gamma_i - \sum_j \gamma_j)^2 = \frac{1}{p_n} \frac{p_n}{2} (\zeta - \frac{1}{2}\zeta)^2 = \frac{1}{8} \zeta^2.$$ 

The above rate is attained and therefore no lower rate exists which bounds all sequences. \hfill \Box

## B.3 Proof of Theorem 3.3 (Sparsity in the limit)

**Proof.** Let us assume that $\lim_{n \to \infty} r = 0$ does not hold. Then, $\exists K_3$: $\forall n \exists m > n$: $|W_m| > K_3 \cdot p_m^{1+\alpha}$. It follows that

$$\frac{1}{p_m} \|C^n\|^2 = \frac{1}{p_m} \sum_{ij} \mathbb{E}^2[x_i x_j] \geq \frac{1}{p_m} K_3 p_m^{1+\alpha} \epsilon^2 = K_3 p_m^{1+\alpha} \epsilon^2.$$

From (LW2), we have

$$\frac{1}{p_m} \|\Gamma^n\|^2 = \frac{1}{p_m} \sum_{ij} \mathbb{E}^2 [y_i^*]^2 \leq \frac{1}{p_m} \sum_{ij} \mathbb{E} [y_i^*] \leq \sqrt{\frac{1}{p_m} \sum_{ij} \mathbb{E} [y_i^*]^2} \leq \sqrt{K_2}.$$

Combining the two equations yields

$$\sqrt{K_2} \geq K_3 \epsilon^2 p_m^{1+\alpha},$$

a contradiction in the limit $m > n \to \infty$. \hfill \Box

## B.4 Proof of Theorem 3.4 (LDL behavior of LW-Shrinkage)

Let us start by proving a lemma:

**Lemma B.1.** Under the assumptions of Theorem 3.4, in the LDL,

$$\lambda^* = \left( 1 + \frac{\sum_i \mathbb{E}^2 \left[ C_{ii} - \hat{T}_{ii} \right]}{\sum_{ij} \text{Var}(\hat{S}_{ij}')} \right)^{-1} + o(1) \quad \text{(B.1)}$$

holds.

**Proof.** For this proof, I start by manipulating the analytic formula for $\lambda^*$ (eq. (3.5)) in the eigenbasis:

$$\lambda^* = \frac{\sum_{i,j} \text{Var}(\hat{S}_{ij}') + \sum_i \left\{ \text{Var}(\hat{T}_{ii}) + \mathbb{E}^2 \left[ C_{ii} - \hat{T}_{ii} \right] \right\}}{\sum_{i,j} \text{Var}(\hat{S}_{ij}') + \sum_i \left( \text{Var}(\hat{T}_{ii}) + \mathbb{E}^2 \left[ C_{ii} - \hat{T}_{ii} \right] \right) + \sum_i \mathbb{E}^2 \left[ C_{ii} - \hat{T}_{ii} \right]}^{-1}.$$
Now, in order to prove the lemma, I have to show that

\[
\begin{align*}
\sum_i \text{Cov}(\hat{S}_{ii}, \hat{T}_{ii}) &= \frac{1}{n} \sum_{i,j} \text{Var}(\hat{S}_{ij}) = o(1), \\
\sum_i \text{Var}(\hat{T}_{ii}) &= \frac{1}{n} \sum_{i,j} \text{Var}(\hat{S}_{ij}) = o(1),
\end{align*}
\]

(i)\(\sum_i \text{Cov}(\hat{S}_{ii}, \hat{T}_{ii}) = \sum_i \text{Var}(\hat{T}_{ii} = \frac{1}{pn} \sum_{ij} \text{Cov}(y_{i1}^2, y_{j1}^2). \quad (B.2)\]

holds. For the covariance expression in (i), we have

\[
\begin{align*}
\sum_{ij} \text{Cov}(\hat{S}_{ij}, \hat{T}_{ij}) &= \sum_i \text{Cov}(\hat{S}_{ii}, \hat{T}_{ii}) = \sum_i \text{Cov} \left( \frac{1}{n} \sum_{s} y_{is}^2, \frac{1}{np} \sum_{jt} y_{jt}^2 \right) \\
&= \frac{1}{pn^2} \sum_{ij} \sum_{st} \text{Cov}(y_{i1}^2, y_{j1}^2) = \frac{1}{pn} \sum_{ij} \text{Cov}(y_{i1}^2, y_{j1}^2).
\end{align*}
\]

For the variance of the \(\hat{T}_{ij}\) in (ii), we have

\[
\begin{align*}
\sum_{ij} \text{Var}(\hat{T}_{ij}) &= \sum_i \text{Var}(\hat{T}_{ii}) = \sum_i \text{Var} \left( \frac{1}{np} \sum_{jt} y_{jt}^2 \right) = \frac{1}{pn^2} \sum_{it} \text{Var} \left( \sum_j y_{jt}^2 \right) \\
&= \frac{1}{pn} \text{Var} \left( \sum_j y_{j1}^2 \right) = \frac{1}{pn} \sum_{ij} \text{Cov}(y_{i1}^2, y_{j1}^2).
\end{align*}
\]

In the next step, I derive a formula for \(\sum_{ij} \text{Var}(\hat{S}_{ij})\), the numerator in (i) and (ii):

\[
\begin{align*}
\sum_{ij} \text{Var}(\hat{S}_{ij}) &= \frac{1}{n^2} \sum_{ij} \text{Var} \left( \sum_t y_{it} y_{jt} \right) = \frac{1}{n} \sum_{ij} \text{Var}(y_{i1} y_{j1}) = \frac{1}{n} \sum_{ij} \{ \mathbb{E} [y_{i1}^2 y_{j1}^2] - \mathbb{E}^2 [y_{i1} y_{j1}] \} \\
&= \frac{1}{n} \sum_{ij} \{ \text{Cov}(y_{i1}^2, y_{j1}^2) + \mathbb{E} [y_{i1}^2] \mathbb{E} [y_{j1}^2] - \mathbb{E}^2 [y_{i1} y_{j1}] \} \geq \frac{1}{n} \sum_{ij} \text{Cov}(y_{i1}^2, y_{j1}^2). \quad (B.3)
\end{align*}
\]

Combining eq. (B.2) and eq. (B.3) yields the desired limit behavior:

\[
\frac{\sum_i \text{Cov}(\hat{S}_{ii}, \hat{T}_{ii})}{\sum_{ij} \text{Var}(\hat{S}_{ij})} = \frac{\sum_i \text{Var}(\hat{T}_{ii})}{\sum_{ij} \text{Var}(\hat{S}_{ij})} \leq \frac{1}{p} = o(1)
\]

and eq. (B.1) follows.

With Lemma B.1, I can now prove Theorem 3.4:

**Proof.** In order to prove the statements, I have to find bounds for the variance term \(\sum_{i,j} \text{Var}(\hat{S}_{ij})\) and the eigenvalue dispersion \(\sum_i \mathbb{E}^2 [C_{ii} - \hat{T}_{ii}]\).
Bounds on the dispersion

The bounds on the dispersion are nearly directly given by (A4), because of the fast convergence of the estimate of the average eigenvalue $\hat{T}_{ii} \to p^{-1} \sum_i \gamma_i$:

$$\sum_i \mathbb{E}^2 \left[ C_{ii} - \hat{T}_{ii} \right] \approx \sum_i (\gamma_i - \frac{1}{p} \sum_j \gamma_j)^2 \overset{(A4)}{=} \Theta \left( \max (p, p^{2k}) \right).$$

Bounds on the variance term

The variance term I analyze in the eigenbasis. For the lower bound, I distinguish two cases:

(i) for $k = 1$, we have

$$\sum_{i,j} \text{Var} (\hat{S}_{ij}) \geq \sum_{i,j} \text{Var} (\hat{S}_{ii}) = \frac{1}{n} \sum_i \left\{ \mathbb{E} [y_{i1}^2] - \mathbb{E}^2 [y_{i1}] \right\} \overset{(A6)}{=} \frac{1}{n} \sum_i \beta_4 \mathbb{E}^2 [y_{i1}^2]
= \frac{\beta_4 p}{n} \frac{1}{p} \sum_i \gamma_i^2 \overset{(A4)}{=} \Theta \left( \max (1, p^{2k-1}) \right) = \Theta(p). (B.4)$$

(ii) for the case $k < 1$, we have

$$\sum_{i,j} \text{Var} (\hat{S}_{ij}) = \frac{1}{n} \sum_{i,j} \left\{ \mathbb{E} [y_{i1}^2 y_{j1}] - \mathbb{E}^2 [y_{i1} y_{j1}] \right\} \geq \frac{1}{n} \sum_{i,j} \left\{ \mathbb{E} [y_{i1}^2] \mathbb{E} [y_{j1}^2] - \mathbb{E}^2 [y_{i1} y_{j1}] \right\}
\geq \frac{1}{n} \left( \sum_i \mathbb{E} [y_{i1}^2] \right)^2 - \frac{1}{n} \sum_i \mathbb{E}^2 [y_{i1}^2] \geq \frac{p^2}{n} \left( \frac{1}{p} \sum_i \mathbb{E} [x_{i1}^2] \right)^2 - \frac{p}{n} \frac{1}{p} \sum_i \gamma_i^2
= \Theta(p) - \Theta(\max(1, p^{2k-1})) = \Theta(p). (B.5)$$

For the upper bound, we have

$$\sum_{i,j} \text{Var} (\hat{S}_{ij}) \leq \frac{1}{n} \sum_{i,j} \left\{ \sqrt{\text{Var}(y_{i1}^2) \text{Var}(y_{j1}^2)} + \mathbb{E} [y_{i1}^2] \mathbb{E} [y_{j1}^2] - \mathbb{E}^2 [y_{i1} y_{j1}] \right\}
\leq \frac{2}{n} \sum_{i,j} \sqrt{\mathbb{E}[y_{i1}^4] \mathbb{E}[y_{j1}^4]} \leq \frac{2p^2}{n} (1 + \alpha_4) \left( \frac{1}{p} \sum_i \mathbb{E}[y_{i1}^2] \right)^2 = \Theta(p). (B.6)$$

As the lower and the upper bound are identical, we have

$$\sum_{i,j} \text{Var} (\hat{S}_{ij}) = \Theta(p).$$

Comparing variance and dispersion limit behavior

Comparing with the dispersion, we see that

- for $k \leq 0.5$ both terms grow at the same rate, therefore $\lambda^*$ neither goes to zero nor goes to one.
- for $k > 0.5$, the dispersion grows with rate $\Theta(p^{2k})$, faster than the rate of the variance. Hence $\lambda^*$ goes to zero.

\[ \square \]

B.5 Proof of Theorem 3.5 (LDL consistency of LW-Shrinkage)

Let us first state some helpful properties of the estimators appearing in $\hat{\lambda}$:
Lemma B.2. \( \text{Var}(\hat{S}_{ij}), \text{Var}(\hat{T}_{ij}), \sum ij \text{Cov}(\hat{S}_{ij}, \hat{T}_{ij}) \) and \( (\hat{S}_{ij} - \hat{T}_{ij})^2 \) are unbiased estimators of \( \text{Var}(\hat{S}_{ij}), \text{Var}(\hat{T}_{ij}), \sum ij \text{Cov}(\hat{S}_{ij}, \hat{T}_{ij}) \) and \( \mathbb{E}[(\hat{S}_{ij} - \hat{T}_{ij})^2] \).

Proof. follows directly by calculation of the expectations. \( \square \)

The proof of consistency under the new assumptions is similar to the proof for under the assumptions (LW1)-(LW3) in (Ledoit & Wolf, 2004). An additional ingredient is the \( k \)-dependent limit behavior from Theorem 3.4, which allows us to find bounds on specific terms which would otherwise not converge to zero in the limit.

Proof. in a first step, let us rewrite the relative error:

\[
m = \left( \frac{\lambda^* - \hat{\lambda}}{\lambda^*} \right)^2 = \left( 1 - \frac{\hat{\lambda}}{\lambda^*} \right)^2
\]

\[
= 1 - \left( \frac{n_p p^{-2} \sum ij \text{Var}(\hat{S}_{ij})}{p^2 \max(1,2k)} \sum ij (\hat{S}_{ij} - \hat{T}_{ij})^2 \right) \cdot \left( \frac{n_p p^{-2} \sum ij \text{Var}(\hat{S}_{ij})}{p^2 \max(1,2k)} \sum ij \mathbb{E}[(\hat{S}_{ij} - \hat{T}_{ij})^2] \right)^{-1}.
\]

Here, pre-factors \( n_p p^{-2} \) and \( p^2 \max(1,2k) \) have been introduced make the corresponding expressions bounded from above and below. Let us now look at the convergence of these normalized estimators of variance and dispersion.

Bound on the Variance of the Dispersion  The dispersion is

\[
\frac{1}{p^2 \max(1,2k)} \sum ij (\hat{S}_{ij} - \hat{T}_{ij})^2 = \frac{1}{p^2 \max(1,2k)} \sum ij \hat{S}_{ij}^2 - 2 \hat{S}_{ij} \hat{T}_{ij} + \hat{T}_{ij}^2. \tag{B.7}
\]

Let us here only analyze the first term on the r.h.s. which has the highest variance, for the other terms the proof is similar but easier:

\[
\frac{1}{p^2 \max(1,2k)} \sum ij \hat{S}_{ij}^2 = \frac{1}{p^2 \max(1,2k)} \sum ij \left( \frac{1}{n} \sum s y_{is} y_{it} \right)^2 = \frac{p^2}{p^2 \max(1,2k) \eta_i^2} \sum st \left( \frac{1}{p} \sum s y_{is} y_{it} \right)^2
\]

\[
= \frac{p^2}{p^2 \max(1,2k) \eta_i^2} \sum s \left( \frac{1}{p} \sum s y_{is}^2 \right)^2 + \frac{p^2}{p^2 \max(1,2k) \eta_i^2} \sum s,t \neq s \left( \frac{1}{p} \sum s y_{is} y_{it} \right)^2. \tag{B.8}
\]

I show that the variance goes to zero by showing that both terms on the right-hand side have zero variance in the limit. For the first term, we have:

\[
\text{Var} \left( \frac{p^2}{p^2 \max(1,2k) \eta_i^2} \sum s \left( \frac{1}{p} \sum s y_{is}^2 \right)^2 \right) \leq \frac{p^4}{p^2 \max(2,4k) \eta_i^3} \mathbb{E} \left[ \left( \frac{1}{p} \sum s y_{is}^4 \right)^4 \right]
\]

\[
= \frac{p^2}{p^2 \max(2,4k) \eta_i^3} \mathbb{E} \left[ \left( \frac{1}{p} \sum s x_{i1}^2 \right)^4 \right] \leq \frac{p^4}{p^2 \max(2,4k) \eta_i^3} \mathbb{E} \left[ \frac{1}{p} \sum s x_{i1}^4 \right]
\]

\[
\leq \frac{p^4}{p^2 \max(2,4k) \eta_i^3} K_2 = O(p^{- \max(1,4k-1)}) = o(1).
\]
Therefore the first term in eq. (B.8) converges to its expectation. Let us now look at the second term:

\[
\text{Var} \left( \frac{p^2}{p_{\text{max}(1,2k)})^2} \sum_{s,t \neq s} \left( \frac{1}{p} \sum_i y_{is}y_{it} \right)^2 \right) = \frac{p^4}{p_{\text{max}(2,4k)})^4} \sum_{s,t \neq s,t' \neq s'} \text{Cov} \left( \left( \frac{1}{p} \sum_i y_{is}y_{it} \right)^2, \left( \frac{1}{p} \sum_i y_{is'}y_{it'} \right)^2 \right). \tag{B.9}
\]

The covariance expression in eq. (B.9) only depends on the cardinality of the intersection, which I denote by \((\{s, t\} \cup \{s', t'\})^\#\) and which can take the values of 0, 1 and 2.

**Case 1: Cardinality equal to zero.** When this cardinality is zero,

\[
(\{s, t\} \cup \{s', t'\})^\# = 0,
\]

there is independence and the covariance is zero as well.

**Case 2: Cardinality equal to one.** For

\[
(\{s, t\} \cup \{s', t'\})^\# = 1,
\]

we have \(\Theta(n^3)\) expressions of the form

\[
\frac{1}{p^4} \text{Cov} \left( \left( \sum_i y_{i1}y_{i2} \right)^2, \left( \sum_i y_{i1}y_{i3} \right)^2 \right) = \frac{1}{p^4} \mathbb{E} \left[ \left( \sum_i y_{i1}y_{i2} \right)^2 \left( \sum_i y_{i1}y_{i3} \right)^2 \right] - \frac{1}{p^2} \mathbb{E}^2 \left[ \left( \sum_i y_{i1}y_{i3} \right)^2 \right]
\]

which are, as both terms are positive,

\[
\leq \max \left( \frac{1}{p^4} \mathbb{E} \left[ \left( \sum_i y_{i1}y_{i2} \right)^2 \left( \sum_i y_{i1}y_{i3} \right)^2 \right], \frac{1}{p^2} \mathbb{E}^2 \left[ \left( \sum_i y_{i1}y_{i2} \right)^2 \right] \right).
\]

For the first term, we have

\[
\frac{1}{p^4} \mathbb{E} \left[ \left( \sum_i y_{i1}y_{i2} \right)^2 \left( \sum_i y_{i1}y_{i3} \right)^2 \right] = \frac{1}{p^2} \sum_{i,j,i',j'} \mathbb{E} \left[ y_{i1}y_{i2}y_{j1}y_{j'} \right] \mathbb{E} \left[ y_{i2}y_{j2} \right] \mathbb{E} \left[ y_{j3}y_{j3} \right] = \frac{1}{p^2} \sum_{i,j} \mathbb{E} \left[ y_{i1}^2 y_{j1}^2 \right] \mathbb{E} \left[ y_{j2}^2 \right] \mathbb{E} \left[ y_{j3}^2 \right] \leq \frac{1}{p^2} \left( \frac{1}{p} \sum_i \mathbb{E} \left[ y_{i1}^2 \right] \mathbb{E} \left[ y_{i2}^2 \right] \right)^2 \leq \frac{1}{p^2} \left( 1 + \alpha_4 \frac{1}{p^2} \mathbb{E}^2 \left[ y_{i1}^2 \right] \right)^2 \leq \frac{1 + \alpha_4}{p^2} \Theta \left( \max(1, p^{4k-2}) \right).
\]

**(A4)**

For the second term, we have

\[
\frac{1}{p^2} \mathbb{E}^2 \left[ \left( \sum_i y_{i1}y_{i2} \right)^2 \right] = \frac{1}{p^2} \left( \sum_{i,j} \mathbb{E}^2 \left[ y_{i1}y_{j1} \right] \right)^2 \leq \frac{1}{p^2} \left( \frac{1}{p} \sum_i \mathbb{E}^2 \left[ y_{i1}^2 \right] \right)^2 \leq \frac{1}{p^2} \Theta \left( \max(1, p^{4k-2}) \right).
\]

**(A4)**

Therefore, we have, combined with the pre-factors in eq. (B.9),

\[
\frac{p^4(4\Theta(n^3)}{p_{\text{max}(2,4k)})^4} \left| \text{Cov} \left( \left( \frac{1}{p} \sum_i y_{is}y_{it} \right)^2, \left( \frac{1}{p} \sum_i y_{is'}y_{it'} \right)^2 \right) \right| = \frac{4p^2}{p_{\text{max}(2,4k)})^2} \Theta \left( \max(1, p^{4k-2}) \right) = \Theta \left( p^{-1} \right),
\]

**(A4)**
B.5. PROOF OF THEOREM 3.5 (LDL CONSISTENCY OF LW-SHRINKAGE)

and I have shown that for the terms with \((\{s, t\} \cup \{s', t'\})^\# = 1\), the variance goes to zero.

**Case 3: Cardinality equal to two.** For \((\{s, t\} \cup \{s', t'\})^\# = 2\),

we get \(\Theta(n^2)\) expressions of the form

\[
\frac{1}{p^4} \sum_{i,j,i',j'} |\text{Cov} (y_{1i}y_{2i'}y_{1j}y_{2j'}, y_{1j}y_{2j}y_{j'}y_{j'})| \leq \frac{1}{p^4} \sum_{i,j,i',j',j''} |\text{Cov} (y_{1i}y_{2i}y_{i'}y_{1j}y_{2j}y_{j'}y_{j''})|.
\]

In this summation, I decompose the set of integers into two disjoint subsets: \(\{1, \ldots, p\} = Q \cup R\), where \(Q\) is the set of distinct integers and \(R\) is the remainder:

\[
\frac{1}{p^4} \sum_{i,j,i',j'} |\text{Cov} (y_{1i}y_{2i}y_{i'}y_{1j}y_{2j}, y_{1j}y_{2j}y_{j'}y_{j'})| + \frac{1}{p^4} \sum_{i,j,i',j'} |\text{Cov} (y_{1i}y_{2i}y_{i'}y_{1j}y_{2j}, y_{1i}y_{2i}y_{j'}y_{j'})|.
\]

For the sum over \(Q\), it is possible to bring this into a form which is zero as a consequence of (A3):

\[
|\text{Cov} (y_{1i}y_{2i}y_{i'}y_{1j}y_{2j}, y_{1j}y_{2j}y_{j'}y_{j'})| = \|E^2 y_{1i}y_{1j}y_{j'}y_{j'} - E [y_{1i}y_{1j}] E^2 y_{j'}y_{j'}| = \|E^2 y_{1i}y_{1j}y_{j'}y_{j'}| = (\text{Cov} (y_{1i}y_{1j}, y_{1j}y_{j'}))^2.
\]

Taking the pre-factors in eq. (B.9), into account, we get

\[
\frac{\Theta(n^{-2})}{p^{\max(2, 4k)}} \sum_{(i,j,i',j') \in Q} |\text{Cov} (y_{1i}y_{2i}y_{i'}y_{1j}y_{2j}, y_{1j}y_{2j}y_{j'}y_{j'})| \leq O(1) \frac{\sum_{(i,j,i',j') \in Q} |\text{Cov} (y_{1i}y_{1j}, y_{1j}y_{j'})|^2}{|Q_p|} = O(1).
\]

For the sum over \(R\), we have

\[
\frac{1}{p^4} \sum_{i,j,i',j'} |\text{Cov} (y_{1i}y_{2i}y_{i'}y_{1j}y_{2j}, y_{1j}y_{2j}y_{j'}y_{j'})| \leq \frac{1}{p^4} \sum_{i,j,i',j'} 2 |\text{Cov} (y_{1i}y_{2i}y_{i'}y_{1j}y_{2j}, y_{1j}y_{2j}y_{j'}y_{j'})| + 4 |\text{Cov} (y_{1i}y_{2i}y_{1j}y_{2j}, y_{1i}y_{2i}y_{j}y_{j'})|
\]

\[
\leq \frac{1}{p^4} \sum_{i,j,i',j'} 2 \sqrt{E[y_{1i}^2] E[y_{1j}^2] E[y_{j'i}^2] E[y_{j'i}^2]} + 4 \sqrt{E[y_{1i}^2] E[y_{1j}^2] E[y_{j'}^2] E[y_{j'}^2]}
\]

\[
\leq \frac{6}{p^4} \sum_{i,j,i',j'} E[y_{1i}^2] E[y_{1j}^2] E[y_{j'i}^2] E[y_{j'i}^2] \leq \frac{6(1 + \alpha_4)}{p} \left( \frac{1}{p} \sum_i E[y_{i}^2] \right) \left( \frac{1}{p} \sum_j E[y_{j}^2] \right)^2
\]

\[
\leq \frac{6(1 + \alpha_4)}{p} \sqrt{K_2} \Theta \left( \max(1, p^{2k-1}) \right).
\]

Together with the pre-factors in eq. (B.9), we obtain

\[
\frac{\Theta(n^{-2})}{p^{\max(2, 4k)}} \sum_{(i,j,i',j') \in R} |\text{Cov} (y_{1i}y_{2i}y_{i'}y_{1j}y_{2j}, y_{1j}y_{2j}y_{j'}y_{j'})| = \frac{\Theta(n^{-2})}{p^{\max(1, 4k-1)}} O \left( \max(1, p^{2k-1}) \right) = o(1).
\]

and I have shown that the variance goes to zero for the terms with \((\{s, t\} \cup \{s', t'\})^\# = 2\).
Bound on the variance of \( np^{-2} \sum_{i,j} \hat{\text{Var}}(\hat{S}_{ij}) \) Let us first rewrite the sample estimate eq. (3.6):

\[
\frac{n}{p^2} \hat{\text{Var}} \left( \frac{1}{(n-1)n} \sum_i \left( y_{is}y_{js} - \frac{1}{n} \sum_{i',s'} y_{is'}y_{js'} \right)^2 \right) = \frac{1}{(n-1)p^2} \left( \sum_i \left( y_{is}^2y_{js}^2 - \frac{1}{n} \sum_{i',s'} y_{isy_{is'}}y_{jsy_{js'}} \right) = \frac{1}{(n-1)p^2} \left( \sum_n y_{in}y_{jn}^2 - n\hat{S}_{ij}^2 \right). \]

Let us first have a look at the second term. Above, in eq. (B.8), the same term appears with a larger pre-factor. Hence, it also holds that \( \Theta(p^{-2}) \sum_{i,j} \hat{S}_{ij}^2 = o(1) \).

For the first term, we have

\[
\text{Var} \left( \frac{1}{(n-1)p^2} \sum_{i,j} y_{in}y_{jn}^2 \right) = \frac{n}{(n-1)^2p^4} \sum_{i'j'j''} \text{Cov} \left( y_{i_1j_1}^2y_{i_2j_2}^2, y_{i_1'j_1'}^2y_{i_2'j_2'}^2 \right) + \frac{n}{(n-1)^2p^4} \sum_{i'j'j''} \text{Cov} \left( y_{i_1j_1}^2y_{i_2j_2}^2, y_{i_1'j_1'}^2y_{i_2'j_2'}^2 \right). \]

A proof without separating terms is possible, but as we need the results later, I again decompose the set of integers. If all integers are distinct, we obtain

\[
\frac{n}{(n-1)^2p^4} \sum_{i'j'j''} \text{Cov} \left( y_{i_1j_1}^2y_{i_2j_2}^2, y_{i_1'j_1'}^2y_{i_2'j_2'}^2 \right) \leq \frac{n}{(n-1)^2p^4} \sum_{i'j'j''} \sqrt{\mathbb{E}[y_{i_1j_1}^4] \mathbb{E}[y_{i_2j_2}^4] \mathbb{E}[y_{i_1'j_1'}^4] \mathbb{E}[y_{i_2'j_2'}^4]} \leq \frac{n}{(n-1)^2p^4} \sum_{i'j'j''} \sqrt[4]{\mathbb{E}[y_{i_1j_1}^8] \mathbb{E}[y_{i_2j_2}^8] \mathbb{E}[y_{i_1'j_1'}^8] \mathbb{E}[y_{i_2'j_2'}^8]} \leq \frac{n}{(n-1)^2p^4} \left( \frac{1}{p} \sum_i \mathbb{E}[y_i^2] \right)^4 \leq \frac{n}{(n-1)^2p^4} K_2 = O(1). \]

For \((i, i', j, j') \in R\) we have

\[
\frac{n}{(n-1)^2p^4} \sum_{i'j'j''} \text{Cov} \left( y_{i_1j_1}^2y_{i_2j_2}^2, y_{i_1'j_1'}^2y_{i_2'j_2'}^2 \right) \leq \frac{6n}{(n-1)^2p^4} \sum_{i,j,i'} \text{Cov} \left( y_{i_1j_1}^2y_{i_2j_2}^2, y_{i_1'j_1'}^2y_{i_2'j_2'}^2 + \text{Cov} \left( y_{i_1j_1}^2y_{i_2j_2}^4, y_{i_1'j_1'}^2y_{i_2'j_2'}^4 \right) \right. \leq \frac{6n}{(n-1)^2p^4} \sum_{i,j,i'} \sqrt[4]{\mathbb{E}[y_{i_1j_1}^8] \mathbb{E}[y_{i_2j_2}^8] \mathbb{E}[y_{i_1'j_1'}^8] \mathbb{E}[y_{i_2'j_2'}^8] \leq \frac{12n}{(n-1)^2p^4} \sum_{i,j',i'} \mathbb{E}[y_{i_1j_1}^2] \mathbb{E}[y_{i_2j_2}^2] \mathbb{E}[y_{i_1'j_1'}^2] \mathbb{E}[y_{i_2'j_2'}^2] \leq \frac{6n}{(n-1)^2p^4} K_2 O \left( \max(1, p^{2k-1}) \right) = o(1). \]

B.6 Proof of Theorem 3.6 (FOLDL behavior of LW-Shrinkage)

Proof. In order to show the behavior of \( \lambda^* \), I again have to find upper and lower bounds for the variance and the squared bias. Let us start from the expressions in the proof of Theorem 3.4. From the lower
bound of the variance for $k = 1$, eq. (B.4), we have

$$
\sum_{i,j} \text{Var}(\hat{S}_{ij}) \geq \frac{\beta_k p}{n} \sum_i \gamma_i^2 = p \cdot \Theta(\max(1, p^{2k-1})) = \Theta(p^2).
$$

For the case $k < 1$, see eq. (B.5), we have

$$
\sum_{i,j} \text{Var}(\hat{S}_{ij}) \geq \frac{p^2}{n} \left( \frac{1}{p} \sum_i \mathbb{E}[\hat{x}_{ii}^2] \right)^2 - \frac{p}{n} \sum_i \gamma_i^2 = p^2 \cdot \Theta(1) - p \cdot \Theta(\max(1, p^{2k-1})) = \Theta(p^2).
$$

For the upper bound, see eq. (B.6), we have

$$
\sum_{i,j} \text{Var}(\hat{S}_{ij}) \leq \frac{2p^2}{n} (1 + \alpha_i) \left( \frac{1}{p} \sum_i \mathbb{E}[y_{ii}^2] \right)^2 = \Theta(p^2).
$$

From the lower and the upper bound, we see that the variance is of order $\Theta(p^2)$ in the large $p$ limit.

For the dispersion, we have by assumption (A4) (see Theorem 3.4)

$$
\sum_i \mathbb{E}^2 \left[ C_{ii} - \hat{T}_{ii} \right] = \Theta \left( \max(p, p^2) \right).
$$

From this, it can be seen that for $k$ smaller than 1, $\lambda^*$ goes to one, as the squared bias grows slower than the variance of the $\hat{S}_{ij}$. For $k$ equal to one, the two expressions are bounded at the same rate, and therefore $\lambda^*$ neither goes to zero nor goes to 1.

**B.7 Proof of Theorem 3.7 (FOLDL consistency of LW-Shrinkage)**

*Proof.* As in the proof of Theorem 3.5, I rewrite the normalized error $m$:

$$
m = \left( \frac{\lambda^* - \hat{\lambda}}{1 - \lambda^*} \right)^2 \left( \frac{1 - \lambda^*}{1 - \hat{\lambda}} \right)^2 = \left( 1 - \frac{1 - \lambda}{1 - \lambda^*} \right)^2.
$$

Indepedently of $k$, the expectations of the expressions $p^{-2} \sum_{i,j} \text{Var}(\hat{S}_{ij})$ and $p^{-2} \sum_{i,j} (\hat{S}_{ij} - \hat{T}_{ij})^2$ are bounded from above and below.

**Bound on the Variance of the Dispersion** As in the proof of Theorem 3.5, I only show convergence for $\sum_{i,j} \hat{S}_{ij}^2$, the expression with the highest variance (see eq. (B.7)). We use the same decomposition as in eq. (B.8):

$$
\frac{1}{p^2} \sum_{i,j} \hat{S}_{ij}^2 = \frac{1}{n^2} \sum_s \left( \frac{1}{p} \sum_i y_{is}^2 \right)^2 + \frac{1}{n^2} \sum_{s,t \neq s} \left( \frac{1}{p} \sum_i y_{is}y_{it} \right)^2.
$$

The FOLDL setting is easier, because the sums over $s$ and $t$ are finite sums. For the terms in the first sum in eq. (B.14), we have

$$
\text{Var} \left( \frac{1}{p} \sum_i y_{is}^2 \right) = \frac{1}{p^4} \sum_{i,j,i',j'} \text{Cov} (y_{i1}^2, y_{i1}^2, y_{j1}^2, y_{j1}^2)
$$
For the sum over $Q$, we need assumption (A3'):

$$\sum_{i,j,i',j' \in Q} \text{Cov} \left( y_{i1}^2 y_{j1}^2, y_{i'1}^2 y_{j'1}^2 \right) \leq 24 \frac{\sum_{i,j,i',j' \in Q} \text{Cov} \left( y_{i1}^2 y_{j1}^2, y_{i'1}^2 y_{j'1}^2 \right)}{|Q_p|} \equiv o(1).$$

For the sum over $R$, we have, following the same derivation as in eq. (B.12),

$$\sum_{i,j,i',j' \in R} \text{Cov} \left( y_{i1}^2 y_{j1}^2, y_{i'1}^2 y_{j'1}^2 \right) = p^{-4} O(\max(1, p^{2k-1})�).$$

which is $o(1)$ for $k < 1$. For the terms in the second sum in eq. (B.14), we have

$$\text{Var} \left( \left( \frac{1}{p} \sum_i y_{i1} y_{i2} \right)^2 \right) = \frac{1}{p^4} \sum_{i,j,i',j'} \text{Cov} \left( y_{i1} y_{i2} y_{j1} y_{j2}, y_{i'1} y_{i'2} y_{j'1} y_{j'2} \right) \leq \frac{1}{p^4} \left| \sum_{i,j,i',j'} \text{Cov} \left( y_{i1} y_{i'1}, y_{j1} y_{j'1} \right) \right|.$$  \hspace{1cm} (B.16)

For the sum over $Q$, I simplify using (B.10), this yields

$$\frac{1}{p^4} \sum_{i,j,i',j' \in Q} \left| \text{Cov} \left( y_{i1} y_{i2} y_{i'1} y_{i'2}, y_{j1} y_{j2} y_{j'1} y_{j'2} \right) \right| = \frac{1}{p^4} \sum_{i,j,i',j' \in Q} \left( \text{Cov} \left( y_{i1} y_{i'1}, y_{j1} y_{j'1} \right) \right)^2 \leq 24 \sum_{(i,j,i',j') \in Q} \frac{\left( \text{Cov} \left( y_{i1} y_{i'1}, y_{j1} y_{j'1} \right) \right)^2}{|Q_p|} \equiv o(1).$$

For the sum over $R$, we have, as in eq. (B.11),

$$\frac{1}{p^4} \sum_{i,j,i',j' \in R} \left| \text{Cov} \left( y_{i1} y_{i2} y_{i'1} y_{i'2}, y_{j1} y_{j2} y_{j'1} y_{j'2} \right) \right| \leq \frac{1}{p} O(\max(1, p^{2k-1})�),$$

which is $o(1)$ for $k < 1$.

**Bound on the variance of $p^{-2} \sum_{ij} \text{Var} \left( \hat{S}_{ij} \right)$** Let us now look at the numerator in eq. (B.13):

$$\frac{1}{p^2} \sum_{ij} \text{Var} \left( \hat{S}_{ij} \right) = \frac{1}{p^2} \sum_{ij} \left( \frac{1}{(n-1)n} \sum_s \left( y_{is} y_{js} - \frac{1}{n} \sum_{s'} y_{is} y_{js'} \right)^2 \right)$$

$$= \frac{1}{(n-1)n p^2} \sum_{ij} \left( \sum_s y_{is}^2 y_{js}^2 - \frac{1}{n} \sum_{s,s'} y_{is} y_{js} y_{is'} y_{js'} \right)$$

$$= \frac{1}{(n-1)n} \sum_s \left( \frac{1}{p} \sum_{i} y_{is}^2 \right)^2 - \frac{1}{(n-1)p^2} \sum_{ij} \hat{S}_{ij}^2.$$
No convergence for $k = 1$ For $k < 1$, I have shown that the variance is $o(1)$. For $k = 1$, I have only shown that the variance is $O(1)$. It remains to be shown that this bound is tight.

While I can only show that there are cases in which the estimator is not consistent, this holds in general for independent $y_i$. Then, for the sum over $R$ in eq. (B.15) we have

\[
\frac{1}{p^4} \sum_{i,j,i',j' \in R} \text{Cov} \left( y_{i1}^2 y_{j1}^2, y_{i'1}^2 y_{j'1}^2 \right) = \frac{1}{p^4} \sum_{i,j,i',j'} 2\text{Cov} \left( y_{i1}^4, y_{i'1}^2 y_{j1} y_{j'1} \right) + 4\text{Cov} \left( y_{i1}^2 y_{i'1}^2, y_{j1}^2 y_{j'1}^2 \right) \\
\geq \frac{4}{p^4} \sum_{i,j,i',j'} \text{Cov} \left( y_{i1}^2 y_{i'1}^2, y_{j1} y_{j'1} \right) \\
= \frac{4}{p^4} \sum_{i,j,i',j'} \left( E \left[ y_{i1}^4 \right] - E^2 \left[ y_{i1}^2 \right] \right) E \left[ y_{j1}^2 \right] E \left[ y_{j'1}^2 \right] \\
+ \sum_{i,j \in Q_2} E \left[ y_{i1}^4 \right] E \left[ y_{j1}^2 \right] - E^2 \left[ y_{i1}^2 \right] E^2 \left[ y_{j1}^2 \right] \\
+ \sum_{i,j \in Q_2} E \left[ y_{i1}^4 \right] E \left[ y_{j1}^2 \right] - E \left[ y_{i1}^2 \right] E \left[ y_{j1}^2 \right] E \left[ y_{j'1}^2 \right] \\
+ \sum_{i} E \left[ y_{i1}^8 \right] - E^2 \left[ y_{i1}^4 \right],
\]

where $Q_2$ and $Q_3$ are the sets of two and three distinct integers, respectively. Each of the four sums is larger than zero. Therefore, it is sufficient to show that the first sums remains finite in the limit:

\[
\frac{4}{p^4} \sum_{i,j,i',j' \in Q_3} \left( E \left[ y_{i1}^4 \right] - E^2 \left[ y_{i1}^2 \right] \right) E \left[ y_{j1}^2 \right] E \left[ y_{j'1}^2 \right] \overset{(B.10)}{\geq} \frac{4\beta_4}{p^4} \sum_{i,j,i',j'} \left( \frac{1}{p^2} \sum_t \gamma_t^2 \right) \left( \frac{1}{p} \sum_j \gamma_j^2 \right)^2 = O(1).
\]

As the covariance between the sums in eq. (B.14) is positive,

\[
\text{Cov} \left( \sum_i \left( \sum_s y_{is}^2 \right)^2, \sum_i \left( \sum_s y_{is} y_{it} \right)^2 \right) = \sum_s \sum_{i \neq i'} \sum_{j \neq j'} \sum_i \left[ y_{is}^2 y_{js} y_{it} y_{j't} y_{i't} y_{i't'} \right] - E \left[ y_{is}^2 y_{js} \right] E \left[ y_{it} y_{j't} y_{i't} y_{i't'} \right] \geq 0.
\]

the variance of the denominator in eq. (B.13) is positive. And because the numerator does not converge to a multiple of the denominator, the variance of $\hat{\lambda}$ does not converge to zero.

□

B.8 Proof of Theorem 3.8 (Convergence of the eigendecomposition)

Proof. I follow the same steps as in Nadler (2008). Let us define

\[
s_u^2 := \frac{1}{n} \sum_{\nu=1}^n (u^\nu)^2, \quad \kappa := \|v_1\|s_u,
\]
\[ \rho_j := \frac{1}{ns_u} \sum_{\nu=1}^{n} u_{j\nu} u_{j\nu}^\nu, \quad \beta_{ij} := \frac{1}{n} \sum_{\nu=1}^{n} \epsilon_\nu^j \epsilon_\nu^i. \]

With these definitions, let us write down the sample covariance matrix in a basis where \( \mathbf{v} \) is collinear to the first basis vector:

\[
\hat{\mathbf{S}}_n = \begin{pmatrix}
\kappa^2 + 2\kappa \rho_1 + \beta_{11} & b_2 & \ldots & b_p \\
b_2 & \beta_{22} & \ldots & \beta_{2p} \\
\vdots & \vdots & \ddots & \vdots \\
b_p & \beta_{p2} & \ldots & \beta_{pp}
\end{pmatrix},
\]

where \( b_i = \kappa \rho_i + \beta_{1j} \). Let us now rotate in the eigenbasis of the \( \beta \)-submatrix:

\[
\mathbf{V} \hat{\mathbf{S}}_n \mathbf{V}^\top = \begin{pmatrix}
\kappa^2 + 2\kappa \rho_1 + \beta_{11} & \hat{b}_2 & \ldots & \hat{b}_p \\
\hat{b}_2 & \gamma_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
\hat{b}_p & 0 & \ldots & \gamma_p
\end{pmatrix},
\]

where \( \hat{b}_i = \kappa \hat{\rho}_i + \hat{\beta}_{1j} \) with

\[
\hat{\rho}_j = \frac{1}{ns_u} \sum_{\nu=1}^{n} u_{j\nu} \hat{\epsilon}_\nu^j, \quad \hat{\beta}_{ij} = \frac{1}{n} \sum_{\nu=1}^{n} \hat{\epsilon}_\nu^i \hat{\epsilon}_\nu^j.
\]

where \( \hat{\epsilon}_\nu^j \) is the noise in the \( \gamma \)-eigenbasis. Matrices like the one in eq. (B.17) are called arrowhead matrices, their characteristic equations have a simple form:

\[
f(\gamma) = (\gamma - \kappa^2 - 2\kappa \rho_1 - \beta_{11}) - \sum_{j=2}^{p} \frac{\hat{b}_j^2}{\gamma - \gamma_j}.
\]

To solve the characteristic equation, I use the Marčenko-Pastur law which yields the density of the eigenvalues \( \gamma_j \) of the submatrix:

\[
f_{\text{MP}}(x) = \frac{1}{2\pi xc} \sqrt{(b - x)(x - a)}, \quad x \in [a, b],
\]

with \( a = (1 - \sqrt{c})^2 \) and with \( b = (1 + \sqrt{c})^2 \).

As the random variable \( \hat{b}_j \) has variance \( \gamma_j (\kappa^2 + 2\kappa \rho_1 + \beta_{11})/n \), I introduce a new random variable \( \eta_j = (\gamma_j (\kappa^2 + 2\kappa \rho_1 + \beta_{11})/n)^{-1/2} \hat{b}_j \). We obtain

\[
\sum_{j=2}^{p} \frac{\hat{b}_j^2}{\gamma - \gamma_j} = (\kappa^2 + 2\kappa \rho_1 + \beta_{11}) \frac{1}{n} - \frac{1}{p - 1} \sum_{j=2}^{p} \frac{\gamma_j \eta_j^2}{\gamma - \gamma_j}.
\]

In the large \( n, p \) limit, we have \( \kappa^2 \to 1 \), \( \rho_1 = \mathcal{O}(1/\sqrt{p}) \to 0 \) and \( \beta_{11} \to \text{Var}(\beta_{11}) = 1 \). The fluctuations in the \( \eta_j^2 \) average out — therefore, we have

\[
\lim_{p,n \to \infty} \sum_{j=2}^{p} \frac{\hat{b}_j^2}{\gamma - \gamma_j} = (\|\mathbf{V}_1\|^2 + 1) \frac{p - 1}{n} \int_{\alpha}^{b} f_{\text{MP}}(x) \frac{x}{\gamma - x} dx.
\]

Plugging this into eq. (B.18) and solving for the \( \gamma_1 \), we obtain

\[
\gamma_1(\alpha) = \alpha + c \frac{\alpha}{\alpha - 1}.
\]
B.9. PROOF OF THEOREM 3.9 (CONSISTENCY OF OC-SHRINKAGE)

For the eigenvectors of the arrowhead matrix, there is also a closed form expression: up to normalization, the eigenvector corresponding to the eigenvalue $\gamma_1$ is given by

$$v_1 = \left(1, \frac{\hat{b}_2}{\gamma_1 - \gamma_2}, \ldots, \frac{\hat{b}_p}{\gamma_1 - \gamma_p}\right)$$

and the overlap is therefore

$$R^2 = \frac{(v_1, e_1)}{\|v_1\|^2} = \frac{1}{\sum_{j=2}^p \hat{b}_j^2 / (\gamma_j - \gamma_j)^2}.$$ 

In the large $n, p$ limit this again turns into an integral,

$$\lim_{p,n\to\infty} R^2 = \frac{1}{1 + p/n \int_a^b f_{\text{MP}}(x)x/(\gamma_1 - x)^2 \, dx}.$$ 

whose evaluation finishes the proof. \qed

B.9 Proof of Theorem 3.9 (Consistency of oc-shrinkage)

Before I start with the proof of the optimality of oc-Shrinkage, I derive the following lemma:

Lemma B.3. In the limit, for the sum over the squared $\varepsilon_i$, we have

$$\lim_{p \to \infty} \sum_{i \geq 2} \varepsilon_i^2 \leq K_6 n^{-k}.$$ 

Proof. By definition, we have

$$\lim_{p \to \infty} \sum_{i \geq 2} \varepsilon_i^2 = \lim_{p \to \infty} \sum_{i \geq 2} |(e_1, v_i)|^2 = \lim_{p \to \infty} \left(1 - |(e_1, v_1)|^2\right).$$

Using Theorem 3.8 –here the noise is even bounded by a constant– we have

$$\leq \lim_{p \to \infty} \left(1 - \frac{np^{2k-1} - 1}{np^{2k-1} + p^k}\right) \leq \lim_{p \to \infty} \left(\frac{p^k}{np^{2k-1} + p^k}\right)$$

$$\leq \lim_{p \to \infty} \left(\frac{1}{np^{k-1} + 1}\right) \leq \lim_{p \to \infty} K_6 n^{-k}.$$ 

Corollary B.1. In the limit, for the sum over the absolute values of the $\varepsilon_i$,

$$\lim_{p \to \infty} \sum_{i \geq 2} |\varepsilon_i| \leq \sqrt{K_6 n^{1/2 - k/2}}$$

holds.

Proof. Let us assume

$$\sum_{i \geq 2} \varepsilon_i^2 = K_6 n^{-k}.$$ 

Then, the sum over the absolute values of $\varepsilon_i$ is maximized if $\varepsilon_i = \varepsilon$ holds. Then, we have
\[
\sum_{i \geq 2} \varepsilon_i^2 = p\varepsilon^2 = K_6n^{-k} \iff \varepsilon = \pm \sqrt{K_6n^{-k/2}}/2
\]
\[
\Rightarrow \sum_{i \geq 2} |\varepsilon_i| \leq \sqrt{K_6n^{1/2-k/2}}.
\]

For Theorems 3.9 and 3.10 we need to know under which circumstances we are allowed to exchange maximization and limit:

**Lemma B.4.** Let $X_1, \ldots, X_p$ be a set of random variables, $\mu_i := \mathbb{E}[X_i]$, $\sigma_i^2 := \mathbb{E}[X_i^2]$ and $\mathbb{E}[X_i^4] < \infty$, and let $x_{im}$ denote the $n^{th}$ realization of the $i^{th}$ random variable. Then, in the LDL,
\[
\lim_{p \to \infty} \max_i \frac{1}{n} \sum_m x_{im} = \max_i \mathbb{E}[X_i] := \mu_{\max}.
\]

**Proof.** It is the sufficient to show that the stronger statement
\[
\lim_{p \to \infty} \max_i \frac{1}{n} \sum_m (x_{im} - \mu_i) = 0
\]
holds. We show that both variance and expectation of the l.h.s. of eq. (B.20) are zero in the limit. Let us first consider the expectation. We require the fact that for the maximum $M$ of Gaussian variables $G_1, \ldots, G_n \sim N(0, \sigma^2)$ we have
\[
\exp(t\mathbb{E}[M]) \leq \mathbb{E}[\exp(M)] \leq \sum_i \mathbb{E}[\exp(tG_i)] = n \exp(t^2 \sigma^2),
\]
where (i) is Jensen’s inequality, (ii) is the union bound, and (iii) follows from the definition of the moment generating function (mgf). This yields
\[
\iff \mathbb{E}[M] \leq \sigma \sqrt{2 \sqrt{\log n}} \quad \text{for} \quad t = \sigma^{-1} \sqrt{2 \log n}.
\]
From the central limit theorem follows that the moment generating function of $\frac{1}{\sqrt{n}} \sum_i (x_{im} - \mu_i)$ converges to the moment generating function (mgf) of a Gaussian, we have
\[
\lim_{p \to \infty} \mathbb{E} \left[ \max_i \frac{1}{n} \sum_m (x_{im} - \mu_i) \right] \leq \lim_{p \to \infty} \frac{\sqrt{2 \log n}}{\sqrt{n}} \max_i \sigma_i = 0.
\]
For the variance, following the same steps yields
\[
\lim_{p \to \infty} \text{Var} \left( \max_i \frac{1}{n} \sum_m (x_{im} - \mu_i) \right) \leq \mathbb{E} \left[ \left( \max_i \frac{1}{n} \sum_m (x_{im} - \mu_i) \right)^2 \right] = 0.
\]
Proof. The normalized squared error on the sample orthogonal complement is defined by
\[ \Delta_{\hat{\omega}}(\lambda) := \frac{1}{p} \sum_{i,j \geq 2} \left( C_{ij}^{\omega} - (1 - \lambda) \tilde{S}_{ij}^{\omega} - \lambda \hat{\nu}_{\hat{\omega}} I_{ij} \right)^2 \]
\[ = \frac{1}{p} \sum_{i,j \geq 2} (C_{ij}^{\omega} - \tilde{S}_{ij}^{\omega})^2 + \lambda^2 (\tilde{S}_{ij}^{\omega} - \hat{\nu}_{\hat{\omega}} I_{ij})^2 + 2\lambda (C_{ij}^{\omega} - \tilde{S}_{ij}^{\omega}) (\tilde{S}_{ij}^{\omega} - \hat{\nu}_{\hat{\omega}} I_{ij}). \]
Minimization yields
\[ \lambda_{\hat{\omega}}^* := \arg \min_{\lambda} \Delta_{\hat{\omega}}(\lambda) = \frac{\frac{1}{p} \sum_{i,j \geq 2} (C_{ij}^{\omega} - \tilde{S}_{ij}^{\omega}) (\tilde{S}_{ij}^{\omega} - \hat{\nu}_{\hat{\omega}} I_{ij})}{\frac{1}{p} \sum_{i,j \geq 2} (\tilde{S}_{ij}^{\omega} - \hat{\nu}_{\hat{\omega}} I_{ij})^2} =: \frac{W_{\hat{\omega}}}{E_{\hat{\omega}}}. \]

We have to show that, in the limit, this is equal to
\[ \hat{\lambda}_{\hat{\omega}} := \frac{\frac{1}{p} \sum_{i,j \geq 2} \text{Var}(\tilde{S}_{ij}^{\omega}) - \text{Cov}(\tilde{S}_{ij}^{\omega}, \hat{\nu}_{\hat{\omega}} I_{ij})}{\frac{1}{p} \sum_{i,j \geq 2} (\tilde{S}_{ij}^{\omega} - \hat{T}_{\hat{\omega},ij})^2} =: \frac{\hat{V}_{\hat{\omega}}}{E_{\hat{\omega}}}. \]

As the denominator is the same, it has to be shown that the numerator is identical in the limit. Therefore, the following equalities have to be proven:
\[ \lim_{p \to \infty} W_{\hat{\omega}} \overset{(1)}{=} \lim_{p \to \infty} W_{oc} \overset{(2)}{=} \lim_{p \to \infty} V_{oc} \overset{(3)}{=} \lim_{p \to \infty} \hat{V}_{oc} \overset{(4)}{=} \lim_{p \to \infty} \hat{V}_{\hat{\omega}}, \]
where the subscript oc denotes quantities on the population orthogonal complement. For example,
\[ W_{oc} := \frac{1}{p} \sum_{i,j \geq 2} \text{Var}(\tilde{S}_{ij}^{\omega}) - \text{Cov}(\tilde{S}_{ij}^{\omega}, \hat{\nu}_{oc} I_{ij}), \]
with \( \nu_{oc} = (p - 1)^{-1} \sum_{i \geq 2} \tilde{S}_{ii}^{\omega}. \) The proof of eq. (B.21) is based on a set of lemmas (see Section B.10.1):
- Equalities (1) and (4) state that, in the limit, it does not matter if the quantities are calculated on the population or on the sample orthogonal complement. Equality (1) is proven in Lemma B.5, equality (4) is proven in Lemma B.7.
- On the population orthogonal complement, it is possible to prove the desired equality of \( V \) and \( W \). Equality (2) is proven in Lemma B.6 and equality (3) holds by the consistency of LW-Shrinkage.

\[ \square \]

B.10 Proof of Theorem 3.10 (Convergence of the error)

Proof. For the convergence of the difference \( \hat{\delta}_{\hat{\omega}}(\lambda, \hat{\nu}, \hat{\lambda}_{\hat{\omega}}, \hat{\nu}_{\hat{\omega}}) - \delta_{\hat{\omega}}(\lambda, \hat{\nu}, \hat{\lambda}_{\hat{\omega}}, \hat{\nu}_{\hat{\omega}}) \), it is sufficient to show that the estimates of the normalized ESE converge to the normalized SE:
\[ p^{-1} \Delta_{\hat{\omega}}(\lambda, \hat{\nu}) \rightarrow p^{-1} \Delta_{\hat{\omega}}(\hat{\lambda}, \hat{\nu}), \]
\[ p^{-1} \Delta_{\hat{\omega}}(\hat{\lambda}_{\hat{\omega}}, \hat{\nu}_{\hat{\omega}}) \rightarrow p^{-1} \Delta_{\hat{\omega}}(\hat{\lambda}_{\hat{\omega}}, \hat{\nu}_{\hat{\omega}}). \]
The proof is based on the fact that the estimation error in the first PCA direction gets irrelevant in the large \( n, p \) limit and consists of showing the following equivalencies:
\[ \lim_{p \to \infty} \Delta_{\hat{\omega}}(\hat{\lambda}, \hat{\nu}) \overset{(1)}{=} \lim_{p \to \infty} \Delta_{oc}(\hat{\lambda}, \hat{\nu}) \overset{(2)}{=} \lim_{p \to \infty} \Delta_{oc}(\hat{\lambda}, \hat{\nu}) \overset{(3)}{=} \lim_{p \to \infty} \Delta_{oc}(\hat{\lambda}, \hat{\nu}) \overset{(4)}{=} \lim_{p \to \infty} \Delta_{\hat{\omega}}(\hat{\lambda}, \hat{\nu}). \]
For equality (1), I have to show
\[ \lim_{p \to \infty} \hat{V}_{oc} = \lim_{p \to \infty} \hat{V}_{oc} \quad \text{and} \quad \lim_{p \to \infty} \hat{E}_{oc} = \lim_{p \to \infty} \hat{E}_{oc}. \]
The left equality follows from Lemma B.7, the second from Lemma B.8.

For equality (2), I have to show
\[ \lim_{p \to \infty} \hat{V}_{oc} = \mathbb{E} \left[ \hat{V}_{oc} \right] \quad \text{and} \quad \lim_{p \to \infty} \hat{E}_{oc} = \mathbb{E} \left[ \hat{E}_{oc} \right], \]
which both have been shown for the consistency of Shrinkage.

For equality (3), I have to show
\[ \mathbb{E} \left[ \hat{V}_{oc} \right] = \lim_{p \to \infty} W_{oc} \quad \text{and} \quad \mathbb{E} \left[ \hat{E}_{oc} \right] = \lim_{p \to \infty} \hat{E}_{oc}. \]
The left equality follows from Lemma B.6, the right one is again from the consistency of LW-Shrinkage.

For equality (4), I have to show
\[ \lim_{p \to \infty} W_{oc} = \lim_{p \to \infty} W_{\hat{c}c} \quad \text{and} \quad \lim_{p \to \infty} \hat{E}_{oc} = \lim_{p \to \infty} \hat{E}_{\hat{c}c}. \]
The left equality is proven in Lemma B.5, the right one again follows from the consistency of LW-Shrinkage.

For Lemmas B.5, B.6, B.7 and B.8, see Section B.10.1.

\[ \square \]

### B.10.1 Additional lemmas

For the following lemmas, I use the rotated version of the data defined in eq. (3.11), using the basis from eq. (3.10).

**Lemma B.5.** Under the above assumptions,
\[ \lim_{p \to \infty} W_{oc} = \lim_{p \to \infty} W_{\hat{c}c}. \]

**Proof of Lemma B.5.** We have
\[ W_{\hat{c}c} = \sum_{i,j \geq 2} \frac{1}{p} (C_{ij}^{\tau} - \hat{S}_{ij}^{\tau}) (\hat{S}_{ij}^{\tau} - \hat{\nu}_{\hat{c}c} I_{ij}). \]
\[ = p^{-1} \sum_{i,j \geq 2} C_{ij}^{\tau} \hat{S}_{ij}^{\tau} - C_{ij}^{\tau} \hat{\nu}_{\hat{c}c} I_{ij} - \hat{S}_{ij}^{\tau - 2} + \hat{S}_{ij}^{\tau} \hat{\nu}_{\hat{c}c} I_{ij}. \tag{B.22} \]

Now I have to show that, in the limit, the four terms are equal to those in the population orthogonal complement. Let us start with the first term in eq. (B.22):
\[ \lim_{p \to \infty} p^{-1} \sum_{i,j \geq 2} C_{ij}^{\tau} \hat{S}_{ij}^{\tau} = \lim_{p \to \infty} (pn)^{-1} \sum_{i,j \geq 2} \mathbb{E}[u_i u_j] \sum_s u_i u_j s \]
\[ = \lim_{p \to \infty} (pn)^{-1} \sum_{i,j \geq 2} \mathbb{E} \left[ \left( y_i \sqrt{1 - \epsilon_i^2} + y_i \varepsilon_i \right) \left( y_j \sqrt{1 - \epsilon_j^2} + y_j \varepsilon_j \right) \right] \]
B.10. PROOF OF THEOREM 3.10 (CONVERGENCE OF THE ERROR)

\[
\sum_s \left( y_i \sqrt{1 - \varepsilon_i^2} + y_i \varepsilon_i \right) \left( y_j \sqrt{1 - \varepsilon_j^2} + y_j \varepsilon_j \right)
\]

\[
= \lim_{p \to \infty} (pn)^{-1} \sum_{i,j \geq 2} \sum_s \left\{ \mathbb{E} [y_i y_j] y_{is} y_{js} (\varepsilon_i^2 \varepsilon_j^2 - \varepsilon_i^2 - \varepsilon_j^2) + \varepsilon_i \varepsilon_j \mathbb{E} [y_i y_j] y_{is} y_{js} \right\}
\]

\[
\leq \lim_{p \to \infty} 2(pn)^{-1} \sum_{i \geq 2} \mathbb{E} [y_i^2] \varepsilon_i^2 \sum_s y_{is}^2
\]

\[
= \lim_{p \to \infty} 2(pn)^{-1} K_6 n^{-k} \max_{i \geq 2} \mathbb{E} [y_i^2] \sum_s y_{is}^2
\]

\[
= \lim_{p \to \infty} 2p^{-1} K_6 n^{-k} \max_{i \geq 2} \mathbb{E} [y_i^2] \mathbb{E} [y_i^2] = 0.
\]

For term (3), we have

\[
\lim_{p \to \infty} (pn)^{-1} \left| \sum_{i,j \geq 2} \sum_s \mathbb{E} [y_i^2] y_{is}^2 y_{js}^2 \mathbb{E} [\varepsilon_i \varepsilon_j] \right|
\]

\[
\leq \lim_{p \to \infty} (pn)^{-1} K_6^2 n^{-2k} \mathbb{E} [y_i^2] \sum_s y_{is}^2
\]

\[
= \lim_{p \to \infty} (pn)^{-1} K_6^2 n^{-2k} \mathbb{E} [y_i^2] \mathbb{E} [y_i^2] = 0.
\]

Term (5) is

\[
\lim_{p \to \infty} \frac{1}{pn} \left| \sum_{i,j \geq 2} \sum_s \left( 2 \mathbb{E} [y_i y_j] y_{is} y_{js} + \mathbb{E} [y_i y_j] y_{is}^2 + \mathbb{E} [y_i^2] y_{is} y_{js} \right) \sqrt{(1 - \varepsilon_i^2)} \sqrt{(1 - \varepsilon_j^2)} \varepsilon_i \varepsilon_j \right|
\]
APPENDIX B. PROOFS AND DERIVATIONS FOR SHRINKAGE AND STRUCTURE

Let us continue with the second term in eq. (B.22):

\[
\begin{align*}
&\leq \lim_{p \to \infty} \frac{1}{pn} \sum_{i,j \geq 2} \left| \sqrt{(1 - \epsilon_i^2)} \sqrt{(1 - \epsilon_j^2)} \varepsilon_i \varepsilon_j \right| \sum_s \left( E[y_i y_j y_i^2 + E[y_i^2] y_i y_j] \right) \\
&\leq \lim_{p \to \infty} \frac{1}{pn} \sum_{i,j \geq 2} \varepsilon_i^2 \left| \sum_s E[y_i^2] y_i y_j \right| + \sum_{i,j \geq 2} \varepsilon_i \varepsilon_j \left| \sum_s E[y_i^2] y_i y_j \right| \\
&\leq \lim_{p \to \infty} \frac{1}{pn} \left\{ \sum_{i \geq 2} K_6 n^{-k} \left( E[y_i^2] \max_{i \geq 2} E[y_i^2] + E[y_i^2] \max_{i \geq 2} E[y_i^2] \right) \right. \\
&\quad + 2E[y_i^2] n^{-1} \sum_{i,j \geq 2, j \neq i} \varepsilon_i^2 \left| \sum_s y_i y_j \right| \left\} \\
&\leq \lim_{p \to \infty} \frac{2K_6}{pn} \left\{ E[y_i^2] \max_{i \geq 2} E[y_i^2] + E[y_i^2] \max_{i \geq 2, j \neq i} \sum s E[y_i y_j] \right\} = 0.
\end{align*}
\]

For the term (6), we have

\[
\begin{align*}
&\leq \lim_{p \to \infty} \frac{1}{pn} \sum_{i,j \geq 2} \sum_s 2E[y_i y_j] y_i y_j \left| y_i y_j \right| \left| (1 - \epsilon_i^2) \sqrt{(1 - \epsilon_j^2)} \varepsilon_j \right| \\
&\leq \lim_{p \to \infty} 2p^{-1} \sum_{i \geq 2} \left| \varepsilon_i \right| \sum_s \left| \sum_j \left| y_i y_j \right| \right| \\
&\leq \lim_{p \to \infty} 2p^{-1} n^{-1/2} \sum_{i \geq 2} \left| \sum_j \left| y_i \right| \right| \eta_i \eta_1 \\
&\leq \lim_{p \to \infty} 2p^{-1} n^{-1/2} \sum_{i \geq 2} \left| E[y_i] E[\eta_i \eta_1] \right| = 0.
\end{align*}
\]

Finally, we have term (9)

\[
\begin{align*}
&\leq \lim_{p \to \infty} \frac{1}{pn} \sum_{i,j \geq 2} \sum_s 2E[y_i^2] y_i y_j \sqrt{(1 - \epsilon_i^2) \varepsilon_i^2} \\
&\leq \lim_{p \to \infty} 2p^{-1} K_6 n^{k/2} E[y_i^2] \sum_{i \geq 2} \left| \epsilon_i \right| \sum_s \xi_1 \eta_1 \\
&\leq \lim_{p \to \infty} 2p^{-1} K_6 n^{k/2} n^{1/2} E[y_i^2] \max_{i \geq 2} \left| \sum_s \eta_i \eta_1 \right| \\
&= \lim_{p \to \infty} 2p^{-1} K_6 n^{k/2} n^{1/2} E[y_i^2] \max_{i \geq 2} \left| E[\eta_i \eta_1] \right| = 0.
\end{align*}
\]

Let us continue with the second term in eq. (B.22):

\[
\begin{align*}
&\lim_{p \to \infty} p^{-1} \sum_{i,j \geq 2} C_{ij}^\tau \bar{v}_{ij} = \lim_{p \to \infty} p^{-1} \sum_{i,j \geq 2} C_{ij}^\tau \bar{v}_{ij}
\end{align*}
\]
The first term is equal to $\Gamma_{ij} \widehat{S}_{ij}$, I therefore have to show that the remainder goes to zero:

$$
\lim_{p \to \infty} \left| p^{-2} n^{-1} \sum_{i,j \geq 2} \widehat{S}_{ij} \right| \\
\leq \lim_{p \to \infty} p^{-2} K_6 n^{-k} \max_{i \geq 2} |E[y_i^2] + E[y_i^2]| \\
\max_{k \geq 2} n^{-1} \sum_{s} \left( -y_{ks}^2 K_6 n^{-k} + y_{ks}^2 K_6 n^{-k} + 2y_{ks} y_{is} \sqrt{K_6 n^{-k/2}} \right) \\
= \lim_{p \to \infty} p^{-2} K_6 n^{-2k} \max_{i \geq 2} |E[y_i^2] + E[y_i^2]| \max_{k \geq 2} \left| -E[y_k^2] + E[y_k^2] \right| = 0.
$$

Let us now have a look at the third term in eq. (B.22):

$$
\lim_{p \to \infty} p^{-1} \sum_{i,j \geq 2} \widehat{S}_{ij}^2 = \lim_{p \to \infty} p^{-1} n^{-2} \sum_{i,j \geq 2} u_{is} u_{js} \sum_{t} u_{it} u_{jt} \\
= \lim_{p \to \infty} p^{-1} n^{-2} \sum_{i,j \geq 2} \sum_{s,t} \left( y_{js} \sqrt{1 - \varepsilon_i^2} + y_{is} \varepsilon_j \right) \left( y_{js} \sqrt{1 - \varepsilon_i^2} + y_{is} \varepsilon_j \right) \\
\cdot \left( y_{jt} \sqrt{1 - \varepsilon_j^2} + y_{jt} \varepsilon_i \right) \left( y_{jt} \sqrt{1 - \varepsilon_j^2} + y_{jt} \varepsilon_i \right) \\
= \lim_{p \to \infty} p^{-1} n^{-2} \sum_{i,j \geq 2} \sum_{s,t} \left\{ y_{is} y_{js} y_{it} y_{jt} (\varepsilon_i^2 \varepsilon_j^2 - \varepsilon_i^2 - \varepsilon_j^2) \right. \\
+ y_{it}^2 \varepsilon_i^2 \varepsilon_j^2 + y_{jt}^2 \varepsilon_i^2 \varepsilon_j^2 + y_{is} y_{js} y_{it} y_{jt} (\varepsilon_i^2 \varepsilon_j^2 - \varepsilon_i^2 - \varepsilon_j^2) \\
+ 2y_{it} y_{jt} y_{is} y_{is} (1 - \varepsilon_i^2) \varepsilon_j^2 \\
+ (2y_{it} y_{jt} y_{js} y_{is} + 2y_{is} y_{is} y_{it} y_{jt}) \sqrt{(1 - \varepsilon_i^2)} \sqrt{(1 - \varepsilon_j^2)} \varepsilon_i \varepsilon_j \\
+ 4y_{it} y_{jt} y_{is} y_{is} (1 - \varepsilon_i^2) \sqrt{(1 - \varepsilon_j^2)} \varepsilon_j \\
+ 4y_{it} y_{jt} y_{is} \sqrt{(1 - \varepsilon_i^2)} \varepsilon_i \varepsilon_j \right\}.
$$

The first term is equal to $\widehat{S}_{ij}^2$, I therefore have to show that the terms (2)-(9) go to zero. Let us start with
term (2):

\[
\lim_{p \to \infty} \left| \frac{1}{pm^2} \sum_{i \geq 2, j \geq 2} \sum_{s, t} y_{is} y_{it} y_{js} y_{jt} (\varepsilon_i^2 \varepsilon_j^2 - \varepsilon_i^2 - \varepsilon_j^2) \right| \leq \lim_{p \to \infty} \frac{2}{pm^2} \sum_{i, j \geq 2} \sum_{s, t} y_{is} y_{it} y_{js} y_{jt} \varepsilon_i^2 \\
\leq \lim_{p \to \infty} \frac{2}{pm^2} \sum_{i \geq 2} \sum_{s, t, j \geq 2} y_{is} y_{it} y_{js} y_{jt} \\
 \quad j = i: \sum_{s} y_{is}^4 + \sum_{s, t \neq i} y_{is}^2 y_{it}^2 \\
 \quad j \neq i: \sum_{s, t \neq i} y_{is}^2 y_{it}^2 + \sum_{s, t, j \geq 2} y_{is} y_{it} y_{js} y_{jt} \\
\leq \lim_{p \to \infty} 2K_6 \max_{k \geq 2} \frac{1}{n^{1+k}} \sum_{s, t} \left( \frac{1}{p} E[y_{i1}^4] + \frac{n-1}{p} E^2[y_{i1}^2] + \frac{1}{p} \sum_{j \neq i} E[y_{i1}^2] E[y_{j1}^2] + \frac{n-1}{p} \sum_{j \neq i} E^2[y_{i1} y_{j1}] \right) = 0.
\]

For term (3), we have

\[
\lim_{p \to \infty} \left| \frac{1}{pm^2} \sum_{i \geq 2, j \geq 2} \sum_{s, t} y_{is}^2 y_{it} y_{js}^2 \varepsilon_j^2 \right| \leq \lim_{p \to \infty} \frac{K_6 n^{-2k}}{pm^2} \sum_{s} y_{is}^4 + \sum_{s, t \neq i} y_{is}^2 y_{it}^2 \\
\leq \lim_{p \to \infty} \frac{K_6 n^{-2k}}{p} \left( \frac{1}{n} E[y_{i1}^4] + \frac{n-1}{n} E[y_{i1}^2] E[y_{i1}] \right) \\
\leq \lim_{p \to \infty} \frac{K_6}{p} \left( \frac{1}{n} E[y_{i1}^4] + \frac{n-1}{n} E^2[y_{i1}] \right) = 0.
\]

Term (4) goes to zero as

\[
\lim_{p \to \infty} \left| \frac{1}{pm^2} \sum_{i \geq 2, j \geq 2} \sum_{s, t} 2(1 - \varepsilon_i^2 \varepsilon_j^2) y_{is} y_{it} y_{js} y_{jt} \right| \leq \lim_{p \to \infty} \frac{2K_6 n^{-k}}{pm^2} \sum_{i \geq 2} \sum_{s, t} y_{is} y_{it} y_{js} y_{jt} \\
\leq \lim_{p \to \infty} \frac{2K_6 n^{-k}}{p} \left( \frac{1}{n^2} \sum_{s} y_{is}^2 y_{is}^2 + \frac{1}{n^2} \sum_{s, t \neq i} y_{is} y_{it} y_{js} y_{jt} \right) \\
\leq \lim_{p \to \infty} \frac{2K_6 n^{-k}}{p} \left( \frac{1}{n} E[y_{i1} y_{i1}] + \frac{n-1}{n} E^2[y_{i1}] \right) \\
\leq \lim_{p \to \infty} \frac{2K_6}{p} \left( \frac{1}{n} E[y_{i1}^2 y_{i1}] + \frac{n-1}{n} E^2[y_{i1}] \right) = 0.
\]

For term (5), we have

\[
\lim_{p \to \infty} \left| \frac{1}{pm^2} \sum_{i \geq 2, j \geq 2} \sum_{s, t} \varepsilon_i \varepsilon_j \sqrt{1 - \varepsilon_i^2} \sqrt{1 - \varepsilon_j^2} (y_{is} y_{it} y_{jt} y_{js} y_{is} y_{js} y_{it} y_{jt}) \right| \\
\leq \lim_{p \to \infty} \frac{1}{pm^2} \sum_{i \geq 2, j \geq 2} \varepsilon_i \varepsilon_j \left( \sum_{s, t} 2 \cdot (y_{is} y_{it} y_{jt} y_{js} y_{is} y_{js} y_{it} y_{jt}) \right) \\
\leq \lim_{p \to \infty} \frac{2}{pm^2} \sum_{i \geq 2} \varepsilon_i^2 \left( \sum_{s} (2y_{is}^2 y_{is}^2) + \sum_{s, t \neq i} y_{is} y_{it} y_{jt} y_{js} y_{it} y_{jt} \right).
\]
B.10. PROOF OF THEOREM 3.10 (CONVERGENCE OF THE ERROR)

Let us continue with the fourth and last term in eq. (B.22):

Finally, for term (7), we have

Let us continue with the fourth and last term in eq. (B.22):
= \lim_{p \to \infty} p^{-2} n^{-2} \sum_{i \geq 2} \sum_{t} \left( y_{it} \sqrt{1 - \varepsilon_i^2} + y_{1t} \varepsilon_i \right)^2 \sum_{k \geq 2} \sum_{s} \left( y_{ks} \sqrt{1 - \varepsilon_k^2} + y_{1s} \varepsilon_k \right)^2

= \lim_{p \to \infty} p^{-2} n^{-2} \sum_{i \geq 2} \sum_{t} \left( y_{it}^2 (1 - \varepsilon_i^2) + y_{1t}^2 \varepsilon_i^2 + 2 y_{it} y_{1t} \sqrt{1 - \varepsilon_i^2} \varepsilon_i \right)

\cdot \sum_{k \geq 2} \sum_{s} \left( y_{ks}^2 (1 - \varepsilon_k^2) + y_{1s}^2 \varepsilon_k^2 + 2 y_{ks} y_{1s} \sqrt{1 - \varepsilon_k^2} \varepsilon_k \right)

= \lim_{p \to \infty} p^{-2} n^{-2} \sum_{i,k \geq 2} \sum_{s,t} \left\{ y_{it}^2 y_{ks}^2 \varepsilon_i^2 - \varepsilon_i^2 \varepsilon_k^2 \right\}

+ \lim_{p \to \infty} p^{-2} n^{-2} \sum_{i,k \geq 2} \sum_{s,t} \left\{ y_{it}^2 y_{ks}^2 \varepsilon_i^2 + y_{1t}^2 y_{ks}^2 \varepsilon_k^2 \right\}

+ \lim_{p \to \infty} p^{-2} n^{-2} \sum_{i,k \geq 2} \sum_{s,t} \left\{ 4 y_{it} y_{1t} y_{ks} y_{1s} \sqrt{1 - \varepsilon_i^2} \sqrt{1 - \varepsilon_k^2} \varepsilon_i \varepsilon_k \right\}

+ \lim_{p \to \infty} p^{-2} n^{-2} \sum_{i,k \geq 2} \sum_{s,t} \left\{ 2 y_{it}^2 y_{ks}^2 \varepsilon_i^2 (1 - \varepsilon_k^2) \right\}

+ \lim_{p \to \infty} p^{-2} n^{-2} \sum_{i,k \geq 2} \sum_{s,t} \left\{ 4 y_{it}^2 y_{ks} y_{1s} \varepsilon_i \sqrt{1 - \varepsilon_k^2} \varepsilon_k \right\}

+ \lim_{p \to \infty} p^{-2} n^{-2} \sum_{i,k \geq 2} \sum_{s,t} \left\{ 2 y_{it} y_{1t} y_{ks} y_{1s} \sqrt{1 - \varepsilon_i^2} \sqrt{1 - \varepsilon_k^2} \varepsilon_i \varepsilon_k \right\}

The first term is equal to $\bar{S}^2_{i_j, t_j} \hat{\nu}_{1j} I_{i_j}$, I therefore have to show that the remaining terms (2)-(7) go to zero. First note that the terms (3), (4) and (7) already appeared in the third term in eq. (B.22). For term (2), we have

$$\lim_{p \to \infty} p^{-2} n^{-1} \left\{ \sum_{i,k \geq 2} \sum_{s,t} y_{it}^2 y_{ks}^2 \varepsilon_i^2 \varepsilon_k^2 \right\} \leq \lim_{p \to \infty} 2 p^{-2} n^{-2} \sum_{i,k \geq 2} \sum_{s,t} y_{it}^2 y_{ks}^2 \varepsilon_i^2 \varepsilon_k^2$$

$$\leq \lim_{p \to \infty} 2 p^{-2} n^{-2} K_0 n^{-k} \max_{i \geq 2} \left\{ \sum_{k \geq 2} \sum_{s,t} y_{it}^2 y_{ks}^2 \right\}$$

$$\leq \lim_{p \to \infty} 2 p^{-2} n^{-2} K_0 n^{-k} \max_{i \geq 2} \left\{ \sum_{k \geq 2} \sum_{s} y_{1s}^2 y_{ks}^2 + \sum_{s,t \neq s} y_{1t}^2 y_{ks}^2 \right\}$$

$$\leq \lim_{p \to \infty} 2 p^{-1} K_0 n^{-k-1} \max_{i \geq 2} \left\{ n^{-1} \mathbb{E}[y_i^2 y_k^2] + \frac{n-1}{n} \mathbb{E}[y_i^2] \mathbb{E}[y_k^2] \right\} = 0.$$
As we have shown that all terms eq. (B.22) converge to the corresponding terms in the population orthogonal complement, the same holds for $W_{oc}$.

**Lemma B.6.** Under the above assumptions,

\[
\lim_{p \to \infty} W_{oc} = \lim_{p \to \infty} V_{oc}.
\]

**Proof of Lemma B.6.**

\[
\lim_{p \to \infty} W_{oc} = \lim_{p \to \infty} \sum_{i,j \geq 2} \frac{1}{p} (\Gamma_{ij} - \hat{S}_{ij})(\hat{\nu}_{oc} - \hat{\nu}_{oc} I_{ij})
= \lim_{p \to \infty} p^{-1} \sum_{i,j \geq 2} \Gamma_{ij} \hat{S}_{ij} - \Gamma_{ij} \hat{\nu}_{oc} I_{ij} - \hat{S}_{ij}^2 + \hat{S}_{ij} \hat{\nu}_{oc} I_{ij}
= \lim_{p \to \infty} p^{-1} \sum_{i,j \geq 2} \hat{S}_{ij} \hat{\nu}_{oc} I_{ij} - \Gamma_{ij} \hat{\nu}_{oc} I_{ij} - (\hat{S}_{ij}^2 - \Gamma_{ij} \hat{S}_{ij})
= \lim_{p \to \infty} p^{-1} \sum_{i,j \geq 2} \text{Cov}(\hat{S}_{ij}, \hat{\nu}_{oc} I_{ij}) - \text{Var}(\hat{S}_{ij})
= \lim_{p \to \infty} V_{oc}.
\]

As we see, for the lemma to hold I have to show that equality (a) holds. Convergence of the first and third term to their expectations was already necessary for the consistency of LW-Shrinkage. It remains to be shown that the variance of the second and the fourth term goes to zero.

The variance of the second term in the limit is

\[
\lim_{p \to \infty} \text{Var} \left( p^{-1} \sum_{i,j \geq 2} \Gamma_{ij} \hat{\nu}_{oc} I_{ij} \right) = \lim_{p \to \infty} \frac{1}{p^2} \text{Var} \left( \sum_i \Gamma_{ii} \hat{\nu}_{oc} \right)
= \lim_{p \to \infty} \frac{1}{p^2} \sum_{i,j} \Gamma_{ii} \Gamma_{jj} \text{Cov}(\hat{\nu}_{oc}, \hat{\nu}_{oc})
= \lim_{p \to \infty} \frac{1}{p^2} \sum_{i,j} \Gamma_{ii} \Gamma_{jj} \text{Cov} \left( \frac{1}{np} \sum_{k,s} y_{ks}^2, \frac{1}{np} \sum_{l,t} y_{lt}^2 \right)
= \lim_{p \to \infty} \frac{1}{p^4 n} \sum_{i,j,k,l} \Gamma_{ii} \Gamma_{jj} \text{Cov} \left( y_{k}, y_{l} \right) = 0,
\]

and the variance of the fourth term in the limit is

\[
\lim_{p \to \infty} \text{Var} \left( p^{-1} \sum_{i,j} \Gamma_{ij} \hat{S}_{ij} \right) = \lim_{p \to \infty} \frac{1}{p^2} \text{Var} \left( \sum_i \Gamma_{ii} \hat{S}_{ii}^2 \right) = \lim_{p \to \infty} \frac{1}{p^2} \sum_{i,j} \Gamma_{ii} \Gamma_{jj} \text{Cov} \left( \hat{S}_{ii}, \hat{S}_{jj} \right)
\]
\[ = \lim_{p \to \infty} \frac{1}{p^2} \sum_{i,j} \Gamma_{ii} \Gamma_{jj} \text{Cov} \left( \frac{1}{n} \sum_{s} x_{is}, \frac{1}{n} \sum_{j} x_{jt} \right) \]

\[ = \lim_{p \to \infty} \frac{1}{p^2 n} \sum_{i,j} \Gamma_{ii} \Gamma_{jj} \text{Cov}(y_i^2, y_j^2) = 0. \]

**Lemma B.7.** Under the above assumptions,

\[ \lim_{p \to \infty} \hat{V}_{\infty} = \lim_{p \to \infty} \hat{V}_{\infty}. \]

**Proof of Lemma B.7.** Let us now look at the sum over the $\text{Var}(S_{ij}^*)$, the estimated variance of the covariance matrix in the basis defined above:

\[ \frac{1}{p} \sum_{i \geq 2, j \geq 2} \hat{\text{Var}}(S_{ij}^*) = \frac{1}{p(n-1)n} \sum_{i \geq 2, j \geq 2} \sum_{s} \left( \frac{1}{n} \sum_{v} u_{is} u_{js} - \frac{1}{n} \sum_{v} u_{it} u_{jt} \right)^2 \]

\[ = \frac{1}{p(n-1)n} \sum_{i \geq 2, j \geq 2} \left( \sum_{s} u_{is}^2 u_{js}^2 - \frac{1}{n} \sum_{s,t} u_{it} u_{jt} u_{is} u_{js} \right). \]

Note that the second term was already analyzed in the third term of eq. \((B.22)\). I therefore only have to analyze the first term:

\[ \frac{1}{p(n-1)n} \sum_{i \geq 2, j \geq 2} \sum_{t} u_{it}^2 u_{jt}^2 = \frac{1}{p(n-1)n} \sum_{i \geq 2, j \geq 2} \sum_{t} \left( y_{it} \sqrt{1 - \varepsilon_i^2 + y_{it} \varepsilon_i} \right)^2 \left( y_{jt} \sqrt{1 - \varepsilon_j^2 + y_{jt} \varepsilon_j} \right)^2 \]

\[ = \frac{1}{p(n-1)n} \sum_{i \geq 2, j \geq 2} \sum_{t} \left( y_{it}^2 (1 - \varepsilon_i^2) + y_{it} y_{jt} \varepsilon_i \sqrt{1 - \varepsilon_i^2 + y_{it} \varepsilon_i} \right) \]

\[ \cdot \left( y_{jt}^2 (1 - \varepsilon_j^2) + y_{jt} y_{it} \varepsilon_j \sqrt{1 - \varepsilon_j^2 + y_{jt} \varepsilon_j} \right) \]

\[ = \frac{1}{p(n-1)n} \sum_{i \geq 2, j \geq 2} \sum_{t} \left( y_{it}^2 y_{jt}^2 \right)^{(2)} + y_{it}^2 y_{jt}^2 (\varepsilon_i^2 \varepsilon_j^2 - \varepsilon_i^2 - \varepsilon_j^2) \]

\[ + y_{it}^2 \varepsilon_i^2 \varepsilon_j^2 \]

\[ + 4 y_{it} y_{jt} y_{it}^2 \sqrt{1 - \varepsilon_i^2 \varepsilon_j \varepsilon_j^2 \varepsilon_i^2} \]

\[ + 2 y_{it} y_{jt}^2 \varepsilon_i^2 (1 - \varepsilon_j^2) \]

\[ + 4 y_{it}^2 y_{jt} (1 - \varepsilon_i^2) \sqrt{1 - \varepsilon_j^2 \varepsilon_j^2 \varepsilon_j^2} \]

\[ + 4 y_{it}^3 y_{jt}^2 \varepsilon_i^2 \varepsilon_i^2 \sqrt{1 - \varepsilon_j^2 \varepsilon_j^2}. \]

For term (2), we have

\[ \lim_{p \to \infty} \frac{1}{pn(n-1)} \sum_{i \geq 2, j \geq 2} y_{it}^2 y_{jt}^2 (\varepsilon_i^2 \varepsilon_j^2 - \varepsilon_i^2 - \varepsilon_j^2) \]

\[ \leq \lim_{p \to \infty} \frac{1}{pn(n-1)} \sum_{i \geq 2, j \geq 2} \sum_{t} y_{it}^2 y_{jt}^2 (\varepsilon_i^2 \varepsilon_j^2 - \varepsilon_i^2 - \varepsilon_j^2) \]

\[ \leq \lim_{p \to \infty} \frac{1}{pn(n-1)} \sum_{i \geq 2, j \geq 2} \sum_{t} y_{it}^2 y_{jt}^2 (\varepsilon_i^2 \varepsilon_j^2 - \varepsilon_i^2 - \varepsilon_j^2) \]
B.10. PROOF OF THEOREM 3.10 (CONVERGENCE OF THE ERROR)

For term (5), we have

\[
\frac{1}{p(n-1)} \sum_{i \geq 2, j \geq 2} \sum_{t} y_{it}^4 \leq \lim_{p \to \infty} \frac{2K_6n^{-k}}{n(n-1)} \max_{i \geq 2, j \leq 2} \left( \sum_{t} y_{it}^4 + y_{jt}^4 \right)
\]

and for term (6)

\[
\frac{1}{p(n-1)} \sum_{i \geq 2, j \geq 2} \sum_{t} 4y_{it}^2 y_{jt} y_{1t}(1 - \varepsilon_i^2) \leq \lim_{p \to \infty} \frac{4K_6n^{-k}}{n(n-1)} \max_{i \geq 2, j \geq 2} \left( \sum_{t} y_{it}^4 + y_{jt}^4 \right)
\]
The variance of \( b \)

With this, it is shown that the estimate of \( V \)

Finally, for term (7), we have

In order to evaluate the conservative estimate of the performance difference between LW-Shrinkage and \( \text{oc} \)-shrinkage, we need variance estimates for numerator and denominator in the LW-Shrinkage formula, \( \text{oc} \)-shrinkage, and an estimate for correlation between these terms,

Note that in this section, to simplify the notation, all summations start at index 2. For example, \( \sum_{ijkl} \equiv \sum_{ijkl \geq 2} \).

**B.11 Conservative estimate of the oc-shrinkage improvement**

In order to evaluate the conservative estimate of the performance difference between LW-Shrinkage and \( \text{oc} \)-shrinkage, we need variance estimates for numerator and denominator in the LW-Shrinkage formula, \( \hat{\sigma}^2_{\hat{\alpha}} \) and \( \hat{\sigma}^2_{\hat{E}_{\hat{\alpha}}} \), and an estimate for correlation between these terms,

\[
\text{corr} \left( \hat{\sigma}_{\hat{\alpha}}, \hat{\sigma}_{\hat{E}_{\hat{\alpha}}} \right) = \text{Cov} \left( \hat{\sigma}_{\hat{\alpha}}, \hat{\sigma}_{\hat{E}_{\hat{\alpha}}} \right) / (\sigma_{\hat{\alpha}} \sigma_{\hat{E}_{\hat{\alpha}}}).
\]

I first derive expressions only using the i.i.d. assumption, then simplify by assuming normality. The resulting formulas are quite long, but fast to evaluate numerically.

Note that in this section, to simplify the notation, all summations start at index 2. For example, \( \sum_{ijkl} \equiv \sum_{ijkl \geq 2} \).

**B.11.1 Variance of \( \hat{\sigma}_{\hat{\alpha}} \)**

The variance of \( \hat{\sigma}_{\hat{\alpha}} \) is given by

\[
\text{Var}(\hat{\sigma}_{\hat{\alpha}}) = \text{Var} \left( \frac{1}{p} \sum_{i,j} \left\{ \hat{\text{Var}}(\hat{S}_{ij}) - \hat{\text{Cov}}(\hat{S}_{ij}, \hat{T}_{ij}) \right\} \right)
\]
Under the i.i.d. assumption, this expression simplifies:

\[
\begin{align*}
= \frac{1}{p^2} \text{Var} \left( \frac{1}{(n-1)n^2} \sum_{i,j} \left\{ \sum_{t} \frac{u_{it}^2 u_{jt}^2}{n} - \frac{1}{n} \sum_{s,t} u_{it} u_{jt} u_{is} u_{js} \right\} \right) \\
- \frac{1}{p(n-1)n} \sum_{i,j} \left\{ \sum_{t} \frac{u_{it}^2 u_{jt}^2}{n} - \frac{1}{n} \sum_{s,t} u_{is}^2 u_{jt}^2 \right\} \\
= \frac{1}{p^2(n-1)^2 n^2} \text{Var} \left( \sum_{i,j} \left\{ \frac{p-1}{p} \sum_{t} \frac{u_{it}^2 u_{jt}^2}{n} - \frac{1}{n} \sum_{s,t} u_{it} u_{jt} u_{is} u_{js} + \frac{1}{np} \sum_{s,t} u_{is}^2 u_{jt}^2 \right\} \right) \\
+ \frac{1}{n^2} \text{Cov} \left( \sum_{t,s} u_{it} u_{jt} u_{is} u_{js}, \sum_{t',s'} u_{k' t'} u_{l' t'} u_{k' s'} u_{l' s'} \right) \\
+ \frac{1}{n^2 p^2} \text{Cov} \left( \sum_{t,s} u_{it}^2 u_{jt}^2, \sum_{t',s'} u_{k' t'}^2 u_{l' t'}^2 \right) - \frac{2(p-1)}{np} \text{Cov} \left( \sum_{t} u_{it}^2 u_{jt}^2, \sum_{t',s} u_{k t'} u_{l t'} u_{k s} u_{l s} \right) \\
+ \frac{2(p-1)}{np^2} \text{Cov} \left( \sum_{t} u_{it}^2 u_{jt}^2, \sum_{t',s'} u_{k' t'}^2 u_{l' t'}^2 \right) - \frac{2}{n^2} \text{Cov} \left( \sum_{t,s} u_{it} u_{jt} u_{is} u_{js}, \sum_{t',s'} u_{k' t'} u_{l' t'} u_{k' s'} u_{l' s'} \right) \right). \\
\end{align*}
\]

Under the i.i.d. assumption, this expression simplifies:

\[
\begin{align*}
= \frac{1}{p^2(n-1)^2 n^2} \sum_{i,j,k,l} \left\{ \frac{n(p-1)}{p} \text{Cov} (u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2) \\
+ \frac{1}{n^2} \left\{ n \text{Cov} (u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2) + 4n(n-1) \text{Cov} (u_{i1}^2 u_{j1}^2, u_{k1} u_{l1} u_{k2} u_{l2}) \\
+ 2n(n-1) \text{Cov} (u_{i1} u_{j1} u_{k1} u_{l1} u_{k2} u_{l2}) \\
+ 4n(n-1)(n-2) \text{Cov} (u_{i1} u_{j1} u_{k2} u_{l2}, u_{k1} u_{l1} u_{k3} u_{l3}) \right\} \right) \\
+ \frac{1}{n^2 p^2} \left\{ n \text{Cov} (u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2) + 4n(n-1) \text{Cov} (u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2) \\
+ 2n(n-1) \text{Cov} (u_{i1} u_{j1} u_{k1}^2 u_{l1}^2) \\
+ 4n(n-1)(n-2) \text{Cov} (u_{i1} u_{j1} u_{k1}^2 u_{l1}^2) \right\} \\
- \frac{2(p-1)}{np} \left\{ n \cdot \text{Cov} (u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2) + 2n(n-1) \text{Cov} (u_{i1}^2 u_{j1}^2, u_{k1} u_{l1} u_{k2} u_{l2}) \right\} \\
- \frac{2(p-1)}{np^2} \left\{ n \cdot \text{Cov} (u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2) + 2n(n-1) \text{Cov} (u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2) \right\} \\
+ \frac{2}{n^2 p} \left\{ n \text{Cov} (u_{i1} u_{j1} u_{k1}^2 u_{l1}^2) + 2n(n-1) \text{Cov} (u_{i1} u_{j1}^2 u_{k1} u_{l1}^2) \\
+ 2n(n-1) \text{Cov} (u_{i1} u_{j1}^2 u_{k1}^2 u_{l1}^2) + 2n(n-1) \text{Cov} (u_{i1} u_{j1} u_{k2} u_{l2}, u_{k1}^2 u_{l1}^2) \\
+ 4n(n-1)(n-2) \text{Cov} (u_{i1} u_{j1} u_{k2} u_{l2}, u_{k1}^2 u_{l1}^2) \right\} \right) \\
= \frac{1}{p^2(n-1)^2 n^2} \sum_{i,j,k,l} \left\{ \frac{1}{np^2} - n^2 p - 2np^2 + np + 2n + 2p + 1 \frac{\text{Cov} (u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2)}{np^2} \right\}. \\
\end{align*}
\]
\[\begin{align*}
&= \frac{4(n-1)(np-n-p-1)}{np} \text{Cov} \left( u_{i_1}^2 u_{j_1}^2, u_{k_1} u_{l_1} u_{k_2} u_{l_2} \right) \\
&+ \frac{2(n-1)}{n} \text{Cov} \left( u_{i_1} u_{j_1} u_{i_2} u_{j_2}, u_{k_1} u_{l_1} u_{k_2} u_{l_2} \right) \\
&+ \frac{4(n-1)(n-2)}{np} \text{Cov} \left( u_{i_1} u_{j_1} u_{i_2} u_{j_2}, u_{k_1} u_{l_1} u_{k_3} u_{l_3} \right) \\
&+ \frac{4(n-1)(np-n-p-1)}{np^2} \text{Cov} \left( u_{i_1}^2 u_{j_1}^2, u_{k_1}^2 u_{l_2}^2 \right) \\
&+ \frac{2(n-1)}{np^2} \text{Cov} \left( u_{i_1}^2 u_{j_2}^2, u_{k_1} u_{l_2}^2 \right) \\
&+ \frac{4(n-1)(n-2)}{np^2} \text{Cov} \left( u_{i_1}^2 u_{j_2}^2, u_{k_1}^2 u_{l_3}^2 \right) \\
&+ \frac{4(n-1)}{np} \text{Cov} \left( u_{i_1} u_{j_1} u_{i_2} u_{j_2}, u_{k_1}^2 u_{l_2}^2 \right) \\
&+ \frac{8(n-1)(n-2)}{np} \text{Cov} \left( u_{i_1} u_{j_1} u_{i_2} u_{j_2}, u_{k_1}^2 u_{l_3}^2 \right).
\end{align*}\]

For the nine different covariance expressions, Appendix B.11.4 contains evaluations under normality.

### B.11.2 Variance of \( \hat{E}_{\hat{\phi}} \)

For \( \hat{E}_{\hat{\phi}} \), I have to calculate

\[
\text{Var} \left( \hat{E}_{\hat{\phi}} \right) = \text{Var} \left( \frac{1}{p} \sum_{i,j} \left( \hat{S}_{ij}^{\text{tr}} - \hat{\nu}_{ij} \hat{I}_{ij} \right)^2 \right)
\]

\[
= \text{Var} \left( \frac{1}{pn^2} \sum_{i,j} \left\{ \sum_{s,t} u_{is} u_{js} u_{it} u_{jt} - \frac{2}{p} \sum_{k,l} u_{is} u_{jt} u_{kt} u_{lt} + \frac{1}{p^2} \delta_{ij} \sum_{k,l} \sum_{s,t} u_{ks} u_{lt} u_{kt} u_{lt} \right\} \right)
\]

\[
= \text{Var} \left( \frac{1}{pn^2} \sum_{i,j} \left\{ \sum_{s,t} u_{is} u_{js} u_{it} u_{jt} - \frac{1}{p} \sum_{s,t} u_{is}^2 u_{jt}^2 \right\} \right)
\]

\[
= \frac{1}{p^2 n^4} \sum_{i,j,k,l} \left\{ \text{Cov} \left( \sum_{s,t} u_{is} u_{js} u_{it} u_{jt}, \sum_{s,t} u_{is} u_{js} u_{it} u_{jt} \right) - \frac{2}{p} \text{Cov} \left( \sum_{s,t} u_{is} u_{js} u_{it} u_{jt}, \sum_{s,t} u_{is}^2 u_{jt}^2 \right) \right\}
\]

\[
+ \frac{1}{p^2} \text{Cov} \left( \sum_{s,t} u_{is}^2 u_{jt}^2, \sum_{s,t} u_{is}^2 u_{jt}^2 \right)
\]

Again, this expression simplifies under the i.i.d. assumption:

\[
= \frac{1}{p^2 n^4} \sum_{i,j,k,l} \left\{ n \text{Cov} \left( u_{i_1}^2 u_{j_1}^2, u_{k_1}^2 u_{l_1}^2 \right) + 4n(n-1) \text{Cov} \left( u_{i_1}^2 u_{j_1}^2, u_{k_1} u_{l_1} u_{k_2} u_{l_2} \right) \right.
\]

\[
+ 2n(n-1) \text{Cov} \left( u_{i_1} u_{j_1} u_{i_2} u_{j_2}, u_{k_1} u_{l_1} u_{k_2} u_{l_2} \right) + 4n(n-1)(n-2) \text{Cov} \left( u_{i_1} u_{j_1} u_{i_2} u_{j_2}, u_{k_1} u_{l_1} u_{k_3} u_{l_3} \right) \right.
\]

\[
- \frac{2}{p} \left\{ n \text{Cov} \left( u_{i_1}^2 u_{j_1}^2, u_{k_1}^2 u_{l_1}^2 \right) + 2n(n-1) \text{Cov} \left( u_{i_1}^2 u_{j_1}^2, u_{k_1} u_{l_1} u_{k_2} u_{l_2} \right) \right.
\]

\[
+ 2n(n-1) \text{Cov} \left( u_{i_1} u_{j_1} u_{i_2} u_{j_2}, u_{k_1} u_{l_1} u_{k_2} u_{l_2} \right) + 2n(n-1) \text{Cov} \left( u_{i_1} u_{j_1} u_{i_2} u_{j_2}, u_{k_1}^2 u_{l_2}^2 \right) \right\}
\]
Collecting terms, we obtain

\[ \begin{aligned}
&+ 4n(n - 1)(n - 2) \text{Cov} \left( u_1u_1u_2u_2, u_{k_1}^2u_{t_1}^2 \right) \\
&+ \frac{1}{p^2} \left\{ n \text{Cov} \left( u_1^2u_1^2, u_{k_1}^2u_{t_1}^2 \right) + 4n(n - 1) \text{Cov} \left( u_1^2u_1^2, u_{k_1}^2u_{t_2}^2 \right) \\
&+ 2n(n - 1) \text{Cov} \left( u_{k_1}^2u_{t_2}^2, u_{k_1}^2u_{t_2}^2 \right) \\
&+ 4n(n - 1)(n - 2) \text{Cov} \left( u_{k_1}^2u_{t_2}^2, u_{k_1}^2u_{t_3}^2 \right) \right\}. 
\end{aligned} \]

Collecting terms, we obtain

\[
\begin{aligned}
&= \frac{1}{p^2n^3} \sum_{i,j,k,l} \left\{ \frac{(1)}{n} \sum_{s,t} \text{Cov} \left( u_1^2u_2^2, u_{k_1}^2u_{t_1}^2 \right) \\
&\quad - \frac{2n(n - 1)(p - 1)}{p^2} \text{Cov} \left( u_1^2u_2^2, u_{k_1}^2u_{t_1}^2 \right) \\
&\quad + 2n(n - 1) \text{Cov} \left( u_1^2u_2^2, u_{k_1}^2u_{t_2}^2 \right) \\
&\quad + 4n(n - 1)(n - 2) \text{Cov} \left( u_1^2u_2^2, u_{k_1}^2u_{t_3}^2 \right) \right\}. 
\end{aligned}
\]

The same nine covariance expressions show up. Now, only the calculation of the covariance remains.

### B.11.3 Covariance between \( \hat{V}_{\text{oc}} \) and \( \hat{E}_{\text{oc}} \)

For the covariance, we obtain

\[
\text{Cov} \left( \hat{V}_{\text{oc}}, \hat{E}_{\text{oc}} \right) = \text{Cov} \left( \frac{1}{p(n - 1)n} \sum_{i,j} \left\{ \frac{p - 1}{p} \sum_{s,t} u_i^2u_j^2 - \frac{1}{n} \sum_{s,t} u_iu_ju_isu_js + \frac{1}{np} \sum_{s,t} u_isu_js \right\}, \\
\frac{1}{p^2n^2} \sum_{i,j} \left\{ \sum_{s,t} u_isu_jsu_isu_js + \frac{1}{np} \sum_{s,t} u_isu_js \right\} \right). 
\]

\[
= \frac{1}{p^2(n - 1)n^3} \sum_{i,j,k,l} \left\{ \frac{(p - 1)}{p} \text{Cov} \left( \sum_{t} u_i^2u_j^2, \sum_{t} u_{k_1}^2u_{t_1}^2 \right) \\
- \frac{2}{np} \text{Cov} \left( \sum_{t} u_iu_ju_isu_js, \sum_{t} u_{k_1}^2u_{t_2}^2 \right) \right\}. 
\]
As before, this simplifies under the i.i.d. assumption:

\[
= \frac{1}{p^2(n-1)n^3} \sum_{i,j,k,l} \left\{ \frac{(p-1)}{p} \left\{ n \cdot \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2 \right) + 2n(n-1) \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1} u_{l1} u_{k2} u_{l2} \right) \right\} \\
- \frac{(p-1)}{p^2} \left\{ n \cdot \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2 \right) + 2n(n-1) \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l2}^2 \right) \right\} \right. \\
- \frac{1}{n} \left\{ n \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2 \right) + 4n(n-1) \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1} u_{l1} u_{k2} u_{l2} \right) \\
+ 2n(n-1) \text{Cov} \left( u_{i1} u_{j1} u_{j2} u_{j2}, u_{k1} u_{l1} u_{k2} u_{l2} \right) \\
+ 4n(n-1)(n-2) \text{Cov} \left( u_{i1} u_{j1} u_{j2} u_{j2}, u_{k1} u_{l1} u_{k3} u_{l3} \right) \right\} \\
+ \frac{2}{np} \left\{ n \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2 \right) + 2n(n-1) \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1} u_{l1} u_{k2} u_{l2} \right) \\
+ 2n(n-1) \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l2}^2 \right) + 2n(n-1) \text{Cov} \left( u_{i1} u_{j1} u_{j2} u_{j2}, u_{k1}^2 u_{l2}^2 \right) \\
+ 4n(n-1)(n-2) \text{Cov} \left( u_{i1} u_{j1} u_{j2} u_{j2}, u_{k1}^2 u_{l3}^2 \right) \right\} \\
- \frac{1}{np^2} \left\{ n \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2 \right) + 4n(n-1) \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l2}^2 \right) \\
+ 2n(n-1) \text{Cov} \left( u_{i1}^2 u_{j2}^2, u_{k1}^2 u_{l2}^2 \right) \\
+ 4n(n-1)(n-2) \text{Cov} \left( u_{i1}^2 u_{j2}^2, u_{k1}^2 u_{l3}^2 \right) \right\}. 
\]

Collecting terms, we obtain

\[
= \frac{1}{p^2(n-1)n^3} \sum_{i,j,k,l} \left\{ \frac{(1)}{p^2} \left\{ np^2 - 2np - p^2 + n + 2p - 1 \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2 \right) \right\} \right. \\
\left. + \frac{(2)}{p} \left\{ 2(n-1)(np - n - 2p + 2) \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1} u_{l1} u_{k2} u_{l2} \right) \right\} \right. \\
\left. - \frac{(3)}{p} \left\{ 2(n-1) \text{Cov} \left( u_{i1} u_{j1} u_{j2} u_{j2}, u_{k1} u_{l1} u_{k2} u_{l2} \right) \right\} \right. \\
\left. - \frac{(4)}{p^2} \left\{ 4(n-1)(n-2) \text{Cov} \left( u_{i1} u_{j1} u_{j2} u_{j2}, u_{k1} u_{l1} u_{k3} u_{l3} \right) \right\} \right. \\
\left. - \frac{(5)}{p^2} \left\{ 2(n-1)(np - n - 2p + 1) \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l2}^2 \right) \right\} \right. \\
\left. - \frac{(6)}{p^2} \left\{ 2(n-1) \text{Cov} \left( u_{i1}^2 u_{j2}^2, u_{k1}^2 u_{l2}^2 \right) \right\} \right. \\
\left. - \frac{(7)}{p^2} \left\{ 4(n-1)(n-2) \text{Cov} \left( u_{i1}^2 u_{j2}^2, u_{k1}^2 u_{l3}^2 \right) \right\} \right. \\
\left. + \frac{(8)}{p} \left\{ 4(n-1) \text{Cov} \left( u_{i1} u_{j1} u_{j2} u_{j2}, u_{k1}^2 u_{l2}^2 \right) \right\} \right. \\
\left. + \frac{(9)}{p} \left\{ 8(n-1)(n-2) \text{Cov} \left( u_{i1} u_{j1} u_{j2} u_{j2}, u_{k1}^2 u_{l3}^2 \right) \right\}. \right. 
\]

Again, these are the same nine covariance expressions as for the variances.

**B.11.4 Calculation of the appearing covariance expressions**

In the variance and covariance formulas above, only nine different covariance expressions appear. Assuming normality, calculations in the eigenbasis simplify considerably: under normality, uncorrelatedness
implies independence. Note that in order to keep the presentation simple, the multiplicity of terms has been ignored in the case distinctions.

1\textsuperscript{st} covariance expression
\[
\sum_{ijkl} \text{Cov}(u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2) = \sum_{ijkl} [\mathbb{E}(u_{i1}^2 u_{j1}^2 u_{k1}^2 u_{l1}^2) - \mathbb{E}(u_{i1}^2 u_{j1}^2)\mathbb{E}(u_{k1}^2 u_{l1}^2)] \\
\begin{align*}
&= \sum_i \gamma_i^4 - 3\gamma_i^4 = 12 \sum_i \gamma_i^4 \quad (\text{case 1}) \\
&= 4 \sum_{i,j \neq i} \gamma_i^3 \gamma_j^3 - \gamma_i^3 \gamma_j^3 = 48 \sum_{i,j \neq i} \gamma_i^3 \gamma_j^3 \quad (\text{case 2}) \\
&= 0 \\
&= 2 \sum_{i,j \neq i} \gamma_i^2 \gamma_j^2 - \gamma_i^2 \gamma_j^2 = 16 \sum_{i,j \neq i} \gamma_i^2 \gamma_j^2 \quad (\text{case 3}) \\
&= 0
\end{align*}
\]

2\textsuperscript{nd} covariance expression
\[
\sum_{ijkl} \text{Cov}(u_{i1}^2 u_{j1}^2, u_{k1} u_{l1} u_{k2} u_{l2}) = \mathbb{E}(u_{i1}^2 u_{j1}^2 u_{k1} u_{l1} u_{k2} u_{l2}) - \mathbb{E}(u_{i1}^2 u_{j1}^2)\mathbb{E}(u_{k1} u_{l1} u_{k2} u_{l2})
\]
The expression is zero when \( k = l \) does not hold. Four cases remain:
\[
\begin{align*}
&= \sum_i \left( [\mathbb{E}(u_i^2) - \mathbb{E}(u_i^2)\mathbb{E}(y_i^2)]\mathbb{E}(y_i^2) \right) \\
&= \sum_i 15\gamma_i^4 - 3\gamma_i^4 = 12 \sum_i \gamma_i^4 \quad (\text{case 1}) \\
&= 2 \sum_{i,j \neq i} \gamma_i^3 \gamma_j^3 - \gamma_i^3 \gamma_j^3 = 4 \sum_{i,j \neq i} \gamma_i^3 \gamma_j^3 \quad (\text{case 2}) \\
&= 0 \\
&= 0
\end{align*}
\]

3\textsuperscript{rd} covariance expression
\[
\sum_{ijkl} \text{Cov}(u_{i1} u_{j1} u_{i2} u_{j2}, u_{k1} u_{l1} u_{k2} u_{l2}) = \mathbb{E}(u_{i1} u_{j1} u_{i2} u_{j2} u_{k1} u_{l1} u_{k2} u_{l2}) - \mathbb{E}(u_{i1} u_{j1} u_{i2} u_{j2})\mathbb{E}(u_{k1} u_{l1} u_{k2} u_{l2})
\]
The expression is zero if the indices are not paired. Three cases exist:
\[
\begin{align*}
&= \sum_i \mathbb{E}^2[u_i^2] - \mathbb{E}[y_i^2]^2 = \sum_i 9\gamma_i^4 - \gamma_i^4 = 8 \sum_i \gamma_i^4 \quad (\text{case 1}) \\
&= 0 \\
&= 2 \sum_{i,j \neq i} \mathbb{E}(u_i^2)\mathbb{E}(u_j^2) - 1 \sum_{i,j \neq i} \gamma_i^2 \gamma_j^2 \quad (\text{case 3})
\end{align*}
\]
4th covariance expression

\[
\sum_{ijkl} \text{Cov} \left( u_{i1} u_{j1}, u_{k1} u_{l1} \right) = \mathbb{E} \left[ u_{i1} u_{j1} u_{k1} u_{l1} \right] - \mathbb{E} \left[ u_{i1} u_{j1} \right] \mathbb{E} \left[ u_{k1} u_{l1} \right] - \mathbb{E} \left[ u_{i1} u_{j1} u_{k1} u_{l1} \right] \mathbb{E} \left[ u_{k1} u_{l1} \right] \\
= \left( \mathbb{E} \left[ u_{i1} u_{j1} u_{k1} u_{l1} \right] - \mathbb{E} \left[ u_{i1} u_{j1} \right] \mathbb{E} \left[ u_{k1} u_{l1} \right] \right) \mathbb{E} \left[ u_{k1} u_{l1} \right].
\]

The expression is zero, if either \( i = j \) or \( k = l \) do not hold, or the first expectation separates. Therefore, only the case \( i = j = k = l \) remains:

\[
\sum_i \left( \mathbb{E} \left[ u_i^4 \right] - \mathbb{E}^2 \left[ u_i^2 \right] \right) \mathbb{E}^2 \left[ u_i^2 \right] = \sum_i \left( 3 \gamma_i^2 - \gamma_i^2 \right) \gamma_i^2 = 2 \sum_i \gamma_i^4.
\]

5th covariance expression

\[
\sum_{ijkl} \text{Cov} \left( u_{i1}^2 u_{j1}^2, u_{k1}^2 u_{l1}^2 \right) = \mathbb{E} \left[ u_{i1}^2 u_{j1}^2 u_{k1}^2 u_{l1}^2 \right] - \mathbb{E} \left[ u_{i1}^2 u_{j1}^2 \right] \mathbb{E} \left[ u_{k1}^2 u_{l1}^2 \right] \\
= \left( \mathbb{E} \left[ u_{i1}^2 u_{j1}^2 u_{k1}^2 u_{l1}^2 \right] - \mathbb{E} \left[ u_{i1}^2 u_{j1}^2 \right] \mathbb{E} \left[ u_{k1}^2 u_{l1}^2 \right] \right) \mathbb{E} \left[ u_{k1}^2 u_{l1}^2 \right] \\
= \left\{ \begin{array}{ll}
\sum_{i,j} \left( \mathbb{E} \left[ u_i^2 \right] - \mathbb{E} \left[ u_i^2 \right] \mathbb{E} \left[ u_j^2 \right] \right) \mathbb{E} \left[ u_j^2 \right] & i = j \neq k \neq l \\
\sum_i \left( 15 \gamma_i^2 - 12 \gamma_i^2 \right) \gamma_i^2 & i = j = k \neq l \\
2 \cdot \sum_{i,j \neq i,j} \left( \mathbb{E} \left[ u_i^2 \right] \mathbb{E} \left[ u_j^2 \right] - \mathbb{E}^2 \left[ u_i^2 \right] \mathbb{E} \left[ u_j^2 \right] \right) \mathbb{E} \left[ u_j^2 \right] & i = j \neq k \neq l \\
2 \cdot 3 \sum_{i,j \neq i} \gamma_i^2 \gamma_j^2 & i = j = k \neq l \text{ or } i = j \neq k = l \\
\end{array} \right.
\]

6th covariance expression

\[
\sum_{ijkl} \text{Cov} \left( u_{i1}^2 u_{j2}^2, u_{k1}^2 u_{l2}^2 \right) = \mathbb{E} \left[ u_{i1}^2 u_{j2}^2 u_{k1}^2 u_{l2}^2 \right] - \mathbb{E} \left[ u_{i1}^2 u_{j2}^2 \right] \mathbb{E} \left[ u_{k1}^2 u_{l2}^2 \right] \\
= \left( \mathbb{E} \left[ u_{i1}^2 u_{j2}^2 u_{k1}^2 u_{l2}^2 \right] - \mathbb{E} \left[ u_{i1}^2 u_{j2}^2 \right] \mathbb{E} \left[ u_{k1}^2 u_{l2}^2 \right] \right) \mathbb{E} \left[ u_{k1}^2 u_{l2}^2 \right] \\
= \left\{ \begin{array}{ll}
\mathbb{E} \left[ u_{i1}^2 u_{j2}^2 \right] \mathbb{E} \left[ u_{k1}^2 u_{l2}^2 \right] - \mathbb{E} \left[ u_{i1}^2 u_{j2}^2 \right] \mathbb{E} \left[ u_{k1}^2 u_{l2}^2 \right] & i = j = k = l \\
2 \cdot 3 \sum_{i,j \neq i} \gamma_i^2 \gamma_j^2 & i = j \neq k \neq l \\
\end{array} \right.
\]

7th covariance expression

\[
\sum_{ijkl} \text{Cov} \left( u_{i1}^2 u_{j2}^2, u_{k1}^2 u_{l3}^2 \right) = \mathbb{E} \left[ u_{i1}^2 u_{j2}^2 u_{k1}^2 u_{l3}^2 \right] - \mathbb{E} \left[ u_{i1}^2 u_{j2}^2 \right] \mathbb{E} \left[ u_{k1}^2 u_{l3}^2 \right] \\
= \left( \mathbb{E} \left[ u_{i1}^2 u_{j2}^2 u_{k1}^2 u_{l3}^2 \right] - \mathbb{E} \left[ u_{i1}^2 u_{j2}^2 \right] \mathbb{E} \left[ u_{k1}^2 u_{l3}^2 \right] \right) \mathbb{E} \left[ u_{k1}^2 u_{l3}^2 \right] \mathbb{E} \left[ u_{k1}^2 \right] \mathbb{E} \left[ u_{l3}^2 \right] = \sum_{ijl} \gamma_i^2 \gamma_j^2 \gamma_l.
\]

8th covariance expression

\[
\sum_{ijkl} \text{Cov} \left( u_{i1} u_{j1} u_{k1} u_{l2} u_{j2} u_{k2} u_{l2} u_{k2} \right) = \mathbb{E} \left[ u_{i1} u_{j1} u_{k1} u_{l2} u_{j2} u_{k2} u_{l2} u_{k2} \right] - \mathbb{E} \left[ u_{i1} u_{j1} u_{k1} u_{l2} u_{j2} u_{k2} \right] \mathbb{E} \left[ u_{i1} u_{j2} u_{k2} \right] \\
= \left[ \mathbb{E} \left[ u_{i1} u_{j1} u_{k1} u_{l2} u_{j2} u_{k2} u_{l2} u_{k2} \right] - \mathbb{E} \left[ u_{i1} u_{j1} u_{k1} u_{l2} u_{j2} u_{k2} \right] \mathbb{E} \left[ u_{i1} u_{j2} u_{k2} \right] \right] \\
= \left\{ \begin{array}{ll}
\sum \mathbb{E}^2 \left[ u_i^2 \right] - \mathbb{E}^4 \left[ u_i^2 \right] = 8 \sum_i \gamma_i^4 & i = j = k = l \\
\sum_{i,l \neq i} \mathbb{E} \left[ u_i^2 \right] \mathbb{E} \left[ u_i^2 \right] \mathbb{E} \left[ u_l^2 \right] - \mathbb{E}^3 \left[ u_i^2 \right] \mathbb{E} \left[ u_i^2 \right] \mathbb{E} \left[ u_l^2 \right] = 6 \sum_{i,l \neq i} \gamma_i^3 \gamma_l & i = j = k \neq l \text{ or } i = j \neq k = l \\
\end{array} \right.
\]

The expression is only non-zero when \( i = j \) holds. In addition, either \( k = i \) or \( l = i \) has to hold. 

\[
= \left\{ \begin{array}{ll}
\sum \mathbb{E}^2 \left[ u_i^2 \right] - \mathbb{E}^4 \left[ u_i^2 \right] = 8 \sum_i \gamma_i^4 & i = j = k = l \\
\sum_{i,l \neq i} \mathbb{E} \left[ u_i^2 \right] \mathbb{E} \left[ u_i^2 \right] \mathbb{E} \left[ u_l^2 \right] - \mathbb{E}^3 \left[ u_i^2 \right] \mathbb{E} \left[ u_i^2 \right] \mathbb{E} \left[ u_l^2 \right] = 6 \sum_{i,l \neq i} \gamma_i^3 \gamma_l & i = j = k \neq l \text{ or } i = j \neq k = l \end{array} \right.
\]
B.11. CONSERVATIVE ESTIMATE OF THE OC-SHRINKAGE IMPROVEMENT

$9^{th}$ covariance expression

$$\sum_{ijkl} \text{Cov}(u_{i1}u_{j1}u_{j2}, u_{k1}u_{l1}) = \mathbb{E}[u_{i1}u_{j1}u_{j2}u_{k1}u_{l1}] - \mathbb{E}[u_{i1}u_{j1}u_{j2}]\mathbb{E}[u_{k1}u_{l1}]$$

$$= \left(\mathbb{E}[u_{i1}u_{j1}u_{k1}] - \mathbb{E}[u_{i1}u_{j1}]\mathbb{E}[u_{k1}]\right)\mathbb{E}[u_{j2}u_{l2}]\mathbb{E}[u_{l1}^2].$$

The expression is only non-zero if $i = j = k$ holds:

$$= \sum_{i} \left(\mathbb{E}[u_{i}^4] - \mathbb{E}[u_{i}^2]\mathbb{E}[u_{i}^2]\right)\mathbb{E}[u_{i}^2]\mathbb{E}[u_{i}^2]$$

$$= \sum_{ij} (3\gamma_i^2 - \gamma_i^2)\gamma_i\gamma_j = 2\sum_{ij} \gamma_i^3\gamma_j.$$
Appendix C

Proofs for Shrinkage and autocorrelation

C.1 Proof of Theorem 3.11 (FOLDL consistency under AC)

Proof. The first statement eq. (3.17) follows directly from the assumptions (cmp. Bartz & Müller (2013)). The error expressions follow from a decomposition into bias and variance.

Bias of the Sancetta estimator Let us first restate the estimator:

\[ \hat{\gamma}_{ij}^{\text{San}}(s) := \frac{1}{n} \sum_{t=1}^{n-s} (x_{it}x_{jt} - S_{ij}) (x_{i,t+s}x_{j,t+s} - S_{ij}), \quad (C.1) \]

\[ \text{Var}(S_{ij})^{\text{San},b} := \frac{1}{n} \hat{\gamma}_{ij}^{\text{San}}(0) + \frac{2}{n} \sum_{s=1}^{n-1} \kappa(s/b) \hat{\gamma}_{ij}^{\text{San}}(s), \quad b > 0. \quad (C.2) \]

I start by calculating the bias of the autocovariance estimator eq. (C.1):

\[
\mathbb{E} \left[ \hat{\gamma}_{ij}^{\text{San}}(s) \right] = \mathbb{E} \left[ \frac{1}{n} \sum_{t=1}^{n-s} (x_{it}x_{jt} - S_{ij}) (x_{i,t+s}x_{j,t+s} - S_{ij}) \right], \\
= \frac{1}{n} \sum_{t=1}^{n-s} \{ \mathbb{E} [x_{it}x_{jt}(x_{i,t+s}x_{j,t+s} - S_{ij})] - \mathbb{E} [x_{it}x_{jt}S_{ij}] - \mathbb{E} [S_{ij}x_{i,t+s}x_{j,t+s}] + \mathbb{E} [S_{ij}S_{ij}] \}
\]

\[
= \frac{n-s}{n} \mathbb{E} [x_{it}x_{jt}x_{i,t+s}x_{j,t+s}] - \frac{2}{n^2} \sum_{t=1}^{n-s} \sum_{u=1}^{n} \mathbb{E} [x_{it}x_{jt}x_{iu}x_{ju}] + \frac{n-s}{n^3} \sum_{t,u=1}^{n} \mathbb{E} [x_{it}x_{jt}x_{iu}x_{ju}] \\
= \frac{n-s}{n} \text{Cov} [x_{it}x_{jt}, x_{i,t+s}x_{j,t+s}] - \frac{2}{n^2} \sum_{t=1}^{n-s} \sum_{u=1}^{n} \text{Cov} [x_{it}x_{jt}, x_{iu}x_{ju}] + \frac{n-s}{n^3} \sum_{t,u=1}^{n} \text{Cov} [x_{it}x_{jt}, x_{iu}x_{ju}] \\
= \frac{n-s}{n} \text{Cov} [x_{it}x_{jt}, x_{i,t+s}x_{j,t+s}] - \frac{2}{n^2} \sum_{t=1}^{n-s} \sum_{u=1}^{n} \text{Cov} [x_{it}x_{jt}, x_{iu}x_{ju}] + \frac{n-s}{n^3} \sum_{t=1}^{n-s} \sum_{u=1}^{n} \text{Cov} [x_{it}x_{jt}, x_{iu}x_{ju}] \\
= \frac{n-s}{n} \text{Cov} [x_{it}x_{jt}, x_{i,t+s}x_{j,t+s}] - \frac{2}{n^2} \sum_{t=1}^{n-s} \sum_{u=1}^{n} \text{Cov} [x_{it}x_{jt}, x_{iu}x_{ju}] + \frac{n-s}{n^3} \sum_{t=1}^{n-s} \sum_{u=1}^{n} \text{Cov} [x_{it}x_{jt}, x_{iu}x_{ju}].
\]
APPENDIX C. PROOFS FOR SHRINKAGE AND AUTOCORRELATION

\[ + \frac{1}{n^2} \sum_{t,u=1}^{n} \text{Cov} [x_{it}x_{jt}, x_{iu}x_{ju}] - \frac{s}{n^2} \sum_{t,u=1}^{n} \text{Cov} [x_{it}x_{jt}, x_{iu}x_{ju}] \]

\[ = \frac{n-s}{n} \text{Cov} [x_{it}x_{jt}, x_{i,t+s}x_{j,t+s}] - \frac{n+s}{n} \text{Var} [S_{ij}] + \frac{2}{n^2} \sum_{t=n-s}^{n} \sum_{u=1}^{n} \text{Cov} [x_{it}x_{jt}, x_{iu}x_{ju}] . \]

For the bias of the complete variance estimator eq. (C.2), we then have

\[ \text{Var} [\hat{S}_{ij}^{\text{San,b}}] = \frac{1}{n} \text{Var} [x_{it}x_{jt}] + \frac{2}{n} \sum_{s=1}^{n-1} \frac{n-s}{n} \kappa(s/b) \text{Cov} [x_{it}x_{jt}, x_{i,t+s}x_{j,t+s}] \]

\[ - \frac{1}{n} \text{Var} (S_{ij}) \left( 1 + 2 \sum_{s=1}^{n} \frac{n+s}{n} \kappa(s/b) \right) \]

\[ + \frac{4}{n^3} \left( \sum_{s=1}^{n-1} \sum_{t=n-s}^{n} \sum_{u=1}^{n} \text{Cov} [x_{it}x_{jt}, x_{iu}x_{ju}] \kappa(s/b) \right) . \]

Plugging in the truncated kernel, we obtain

\[ \text{Var} [\hat{S}_{ij}^{\text{San,b}}] = \frac{1}{n} \text{Var} [x_{it}x_{jt}] + \frac{2}{n} \sum_{s=1}^{b} \frac{n-s}{n} \text{Cov} [x_{it}x_{jt}, x_{i,t+s}x_{j,t+s}] - \frac{1}{n} \text{Var} (S_{ij}) \left( 1 + 2 \sum_{s=1}^{b} \frac{n+s}{n} \right) \]

\[ + \frac{4}{n^3} \left( \sum_{s=1}^{b} \sum_{t=n-s}^{n} \sum_{u=1}^{n} \text{Cov} [x_{it}x_{jt}, x_{iu}x_{ju}] \right) . \]

I have shown that eq. (3.20) and eq. (3.21) hold.

**Bias of the BC-estimator**  Let us again first restate the estimator:

\[ \hat{\Phi}_{ij}^{BC}(s) := \frac{1}{n} \sum_{t=1}^{n-s} (x_{it}x_{jt}x_{i,t+s}x_{j,t+s} - S_{ij}^2) , \]

\[ \hat{\text{Var}} (S_{ij})^{\text{BC,b}} := \frac{1}{n-1-2b+b(b+1)/n} \left( \hat{\Phi}_{ij}^{BC}(0) + \frac{2}{n^3} \sum_{s=1}^{n-1} \kappa_{TR}(s/b) \hat{\Phi}_{ij}^{BC}(s) \right) . \]
For the bias-corrected estimator, I also start by calculating the bias of the autocovariance estimator eq. (C.3):

$$
\mathbb{E} \left[ \hat{\gamma}_{ij}^{\text{BC}}(s) \right] = \mathbb{E} \left[ \sum_{t=1}^{n-s} \frac{1}{n} \left( x_{it}x_{jt}x_{i_t+s}x_{j_t+s} - S_{ij}^2 \right) \right]
$$

$$
= \mathbb{E} \left[ x_{it}x_{jt}x_{i_t+s}x_{j_t+s} \right] - \mathbb{E} \left[ S_{ij}^2 \right]
$$

$$
= \mathbb{E} \left[ x_{it}x_{jt}x_{i_t+s}x_{j_t+s} \right] - \text{Var} \left[ S_{ij} \right] - \mathbb{E}^2 \left[ \left( x_{it}x_{jt} \right) \right]
$$

$$
= \frac{n-s}{n} \text{Cov} \left[ x_{it}x_{jt}, x_{i_t+s}x_{j_t+s} \right] - \frac{n-s}{n} \text{Var} \left[ S_{ij} \right].
$$

Plugging this into the expression for the complete variance estimator eq. (C.4), we obtain

$$
\mathbb{E} \left[ \hat{\text{Var}}_{ij}^{\text{BC,b}} \right] = \frac{1}{n-1-2b + \frac{b(b+1)}{n}} \mathbb{E} \left[ \hat{\gamma}_{ij}^{\text{BC}}(0) + 2 \sum_{s=1}^{n-1} \kappa_{\text{TR}}(s/b) \hat{\gamma}_{ij}^{\text{BC}}(s) \right]
$$

$$
= \frac{1}{n-1-2b + \frac{b(b+1)}{n}} \left( n \text{Var} \left[ S_{ij} \right] - 2 \sum_{s=0}^{b+1} \frac{n-s}{n} \text{Cov} \left[ x_{it}x_{jt}, x_{i_t+s}x_{j_t+s} \right] \right)
$$

$$
= \text{Var} \left[ S_{ij} \right] - \frac{2}{n-1-2b + \frac{b(b+1)}{n}} \sum_{s=0}^{b+1} \frac{n-s}{n} \text{Cov} \left[ x_{it}x_{jt}, x_{i_t+s}x_{j_t+s} \right].
$$

I have shown that eq. (3.22) holds.

**Bound on the variance** The variance terms are more involved: Let us exemplarily consider the variance of

$$
\frac{1}{p^2} \hat{\text{Var}}_{ij}^{\text{BC,b}} = \frac{1}{p^2(n-1-2b)} \sum_{ij} \left( \hat{\gamma}_{ij}^{\text{BC}}(0) + 2 \sum_{s=1}^{n-1} \kappa_{\text{TR}}(s/b) \hat{\gamma}_{ij}^{\text{BC}}(s) \right).
$$

It is possible to simplify the analysis of the variance of this expression by looking at each $\hat{\gamma}(s)$ separately: there is only a finite number of terms. Derivations are similar for all terms, I here only show the first term:

$$
\text{Var} \left( \frac{1}{p^2(n-1-2b)} \sum_{ij} \hat{\gamma}_{ij}(0) \right) = \text{Var} \left( \frac{1}{p^2(n-1-2b)} \sum_{ij} \left( \sum_s y_{is}^2 y_{js}^2 - \frac{1}{n} \sum_{st} y_{is} y_{js} y_{is} y_{js} \right) \right)
$$

Ignoring constants I again treat each term separately, the first term yields

$$
\text{Var} \left( \frac{1}{p^2} \sum_{ij} y_{ij1}^2 y_{ij1}^2 \right) = \frac{1}{p^4} \sum_{ijkl} \text{Cov} \left( y_{ij1}^2 y_{jk1}^2, y_{ik1}^2 y_{jk1}^2 \right)
$$
\[ = \frac{1}{p^4} \sum_{i,j;i',j' \in Q} \text{Cov} \left( y_{i1}^2 y_{j1}^2, y_{k1}^2 y_{l1}^2 \right) + \frac{1}{p^4} \sum_{i,j;i',j' \in R} \text{Cov} \left( y_{i1}^2 y_{j1}^2, y_{k1}^2 y_{l1}^2 \right). \]

where I decomposed the set of integers into two disjoint subsets: \( \{1, \ldots, p\}^4 = Q \cup R \), where \( Q \) is the set of distinct integers and \( R \) is the remainder. The first term is taken care of by (AC3'). For the second term, \( (i,i',j,j') \in R \), we have

\[
\frac{1}{p^4} \sum_{i,j,i',j'} \text{Cov} \left( y_{i1}^2 y_{j1}^2, y_{i1}^2 y_{j1}^2 \right) \leq \frac{6}{p^4} \sum_{i,j,i'} \sqrt{\text{E}[y_{i1}^4] \text{E}[y_{j1}^4]} \sqrt{\text{E}[y_{i1}^2] \text{E}[y_{j1}^2]} + \sqrt{\text{E}[y_{i1}^4] \text{E}[y_{j1}^4]} \sqrt{\text{E}[y_{i1}^2]}
\]

\[
\leq \frac{6}{p^4} \sum_{i,j,i'} \sqrt{\text{E}[y_{i1}^8] \text{E}[y_{j1}^8]} \sqrt{\text{E}[y_{i1}^6] \text{E}[y_{j1}^6]} + \sqrt{\text{E}[y_{i1}^8] \text{E}[y_{j1}^8]} \sqrt{\text{E}[y_{i1}^6]}
\]

\[
\leq \frac{12(1 + \alpha_8)}{p^4} \sum_{i,j,i'} \text{E}[y_{i1}^8] \text{E}[y_{j1}^8] \text{E}^2[y_{i1}^2] = O \left( \frac{\sum_i \gamma_i^2}{(\sum_i \gamma_i)^2} \right),
\]

where in the last equality I have used that \( \sum_{i=1}^p \gamma_i = \Theta(p) \). The same asymptotic behavior can be derived for the other terms following very similar steps. The bounds in eq. (3.18) and eq. (3.19) follow directly. □
Appendix D

Proofs for Multi-Target Shrinkage

D.1 Proof of Theorem 4.1 (MTS quadratic program)

Proof. I decompose the ESE into bias and variance

\[
\Delta_{\text{MTS}}(\lambda) = \mathbb{E} \left\| \theta - \hat{\theta}_{\text{MTS}}(\lambda) \right\|^2 = \mathbb{E} \left[ \sum_{i=1}^{q} \left( \hat{\theta}_{\text{MTS}}(\lambda) - \theta_i \right)^2 \right]
\]

(D.1)

This can be simplified to

\[
\Delta_{\text{MTS}}(\lambda) = \sum_{i=1}^{q} \left\{ \sum_{j,k=1}^{K} \lambda_j \lambda_k \mathbb{E} \left[ (\hat{T}^j_i - \hat{\theta}_i)(\hat{T}^k_i - \hat{\theta}_i) \right] + 2 \sum_{k=1}^{K} \lambda_k \mathbb{E} \left[ \hat{T}^k_i - \hat{\theta}_i \right] \right\} + 2 \Delta_{\text{qp}}(\lambda) + \text{const.}
\]

(D.2)

Therefore the sets of \( \lambda \) minimizing \( \Delta_{\text{MTS}}(\lambda) \) and \( \Delta_{\text{qp}}(\lambda) \) are identical. \( \square \)

D.2 Proof of Theorem 4.2 (Consistency of MTS)

Proof. From the constraints, it follows directly that

\[
\| \lambda^* \| = O(1),
\]

(D.3)
and from the definition of $A$ and $b$ follows $\forall k: \tau^k_A \geq \tau^k_0$. I first prove (i). We have $\forall k$:

$$\lambda^* \lambda^* = \sum_{k,l=1}^K \lambda_k^* \lambda_l^* \sum_{i=1}^q \left[ \left( \hat{T}_i^k - \hat{\theta}_i \right) \left( \hat{T}_i^l - \hat{\theta}_i \right) \right]^2 \geq \lambda_k^2 \min_{\alpha \in \mathbb{R}_+^q} \sum_{i=1}^q \left( \sum_{l=1}^K \alpha_l (\hat{T}_i^l - \hat{\theta}_i) \right)^2 \overset{(G4)}{=} \lambda_k^2 \Theta \left( p^\alpha_k \right)$$

(D.4)

We then have $\forall k$:

$$\Theta(p^\alpha_k) \overset{(G1)}{=} \Delta \geq \Delta_{MTS}(\lambda^*) \overset{(D.2)}{=} \lambda^* \lambda^* - 2b^T \lambda^* + \sum_i \text{Var}(\hat{\theta}_i)$$

(D.3),(D.4),(D.5)

$$\geq \lambda_k^2 \Theta(p^\alpha_k) + O(p^\alpha_k).$$

Rearranging yields $\lambda_k = O(p^{0.5(\tau_0^2 - \tau^2_k)})$. To prove statement (i) for $\hat{\lambda}_k$, I first define

$$\hat{\Delta}_{MTS}(\lambda) := \lambda^T \hat{A} \lambda - 2\hat{b}^T \lambda + \sum_{i=1}^p \text{Var}(\hat{\theta}_i).$$

Using the result on the limit behavior of $\lambda^*$, we obtain

$$\lambda^* (A - \hat{A}) \lambda^* = \sum_{k,l=1}^K \lambda_k^* \lambda_l^* (A_{kl} - \hat{A}_{kl}) \overset{(G3)}{=} \sum_{k,l=1}^K \lambda_k^* \lambda_l^* o \left( p^{0.5(\tau_0^2 + \tau^2_k)} \right) = o(p^\alpha)$$

(D.6)

This allows us to calculate

$$\Delta_{MTS}(\lambda^*) - \hat{\Delta}_{MTS}(\lambda^*) = \lambda^* (A - \hat{A}) \lambda^* - 2(b - \hat{b})^T \lambda^* \overset{(D.3),(D.6)}{=} o(p^\alpha).$$

(D.7)

In addition, we calculate

$$\hat{\Delta}_{MTS}(\lambda) - \Delta_{MTS}(\lambda) = \hat{\lambda}^T (\hat{A} - A) \lambda - 2(\hat{b} - b)^T \lambda = \sum_k \hat{\lambda}_k^2 o(p^\alpha_k) + o(p^\alpha)$$

(D.8)

Using these equations, we obtain

$$\Theta(p^\alpha) \geq \Delta_{MTS}(\lambda^*) \overset{(D.7)}{=} \hat{\Delta}_{MTS}(\lambda^*) + o(p^\alpha) \geq \hat{\Delta}_{MTS}(\lambda) + o(p^\alpha)$$

(D.8)

$$\Delta_{MTS}(\hat{\lambda}) \overset{(D.4),(D.5)}{\geq} \hat{\lambda}_k^2 \Theta(p^\alpha_k) + O(p^\alpha_k) + \sum_k \hat{\lambda}_k^2 o(p^\alpha_k).$$

Rearranging yields $\hat{\lambda}_k = O(p^{0.5(\tau_0^2 - \tau^2_k)})$ which concludes (i). To prove statement (ii) we have to relate the difference in ESE to the difference in the estimate of the ESE:

$$\left( \Delta_{MTS}(\lambda) - \Delta_{MTS}(\lambda^*) \right) - \left( \hat{\Delta}_{MTS}(\lambda) - \hat{\Delta}_{MTS}(\lambda^*) \right)$$

$$= \left( \Delta_{MTS}(\lambda) - \hat{\Delta}_{MTS}(\lambda)_{MTS} \right) - \left( \Delta_{MTS}(\lambda^*) - \hat{\Delta}_{MTS}(\lambda^*) \right) \overset{(D.7),(D.8),(i)}{=} o(p^\alpha)$$
Using this and the optimality of $\mathbf{x}^*$ and $\hat{\lambda}$ for $\Delta_{\text{MTS}}(\mathbf{\lambda})$ and $\hat{\Delta}_{\text{MTS}}(\mathbf{\lambda})$, respectively, we obtain

$$0 \leq (\Delta^b)^{-1} \left( \Delta_{\text{MTS}}(\hat{\lambda}) - \Delta_{\text{MTS}}(\mathbf{x}^*) \right) = \Theta(p^{-r_\phi}) \left( \hat{\Delta}_{\text{MTS}}(\hat{\lambda}) - \hat{\Delta}_{\text{MTS}}(\mathbf{x}^*) + o(p^{-r_\phi}) \right) \leq 0 + o(1)$$

which concludes the proof of (ii).

The proof of part (iii) is similar to the one of Theorem 2.1 from (Daniel, 1973). On the convex set we have

$$0 \leq (\hat{\lambda} - \mathbf{x}^*)^T \nabla \Delta_{\text{MTS}}(\mathbf{x}^*)$$
$$0 \leq (\mathbf{x}^* - \hat{\lambda})^T \nabla \hat{\Delta}_{\text{MTS}}(\hat{\lambda})$$

where the gradients are $\nabla \Delta_{\text{MTS}}(\mathbf{\lambda}) = (A\lambda + b)$ and $\nabla \hat{\Delta}_{\text{MTS}}(\mathbf{\lambda}) = (\hat{A}\lambda + \hat{b})$. Multiplying eq. (D.10) by minus one and combining the two equations, we obtain

$$(\hat{\lambda} - \mathbf{x}^*)^T \hat{\nabla} \Delta_{\text{MTS}}(\hat{\lambda}) \leq (\hat{\lambda} - \mathbf{x}^*)^T \nabla \Delta_{\text{MTS}}(\mathbf{x}^*) (D.9)$$

Subtracting $(\hat{\lambda} - \mathbf{x}^*)^T \hat{\nabla} \Delta_{\text{MTS}}(\hat{\lambda})$ from both sides, we obtain

$$(\hat{\lambda} - \mathbf{x}^*)^T \left( \nabla \Delta_{\text{MTS}}(\hat{\lambda}) - \nabla \Delta_{\text{MTS}}(\mathbf{x}^*) \right) \leq (\hat{\lambda} - \mathbf{x}^*)^T \left( \nabla \Delta_{\text{MTS}}(\mathbf{x}^*) - \nabla \hat{\Delta}_{\text{MTS}}(\mathbf{x}^*) \right).$$

The left hand side is

$$(\hat{\lambda} - \mathbf{x}^*)^T \hat{\nabla} (\hat{\lambda} - \mathbf{x}^*) \geq ||\hat{\lambda} - \mathbf{x}^*||^2 \min_{||\alpha||=1} \alpha^T \hat{A} \alpha + (\hat{\lambda} - \mathbf{x}^*)^T (\hat{A} - A)(\hat{\lambda} - \mathbf{x}^*) \stackrel{(G4)}{=} \min_{\alpha \in \mathbb{R}^K} ||\hat{\lambda} - \mathbf{x}^*||^2 \cdot \Theta(p^{-r_\phi}).$$

The right hand side is

$$\left((\hat{\lambda} - \mathbf{x}^*)^T (\hat{A} - A) \mathbf{x}^* + (\hat{\lambda} - \mathbf{x}^*)^T (b - \hat{b})\right) = o(p^{-r_\phi}).$$

by (G1), (G2), (G3) and the rates of the $\lambda_k$ given by (i). Therefore, rearranging yields $||\hat{\lambda} - \mathbf{x}^*||^2 = o(1)$. 

D.3 Proof of Theorem 4.3 (LDL consistency of MTS of the mean)

Proof. Without loss of generality, I assume $\mu = 0$. I start by analyzing the asymptotic behavior of the $\Delta^b$, $A_{kk}$ and $b$, then I prove the consistency of $\hat{A}_{kl}$ and $\hat{b}$.

(G1) & (G2): Asymptotic behavior of $\Delta^b$, $A_{kk}$ and $b$ I start with the asymptotic behavior of $\Delta^b = b$. We have

$$\Delta^b = b = \sum_{i=1}^{p} \text{Var}(\hat{\mu}_i) = n^{-1} \sum_{i=1}^{p} \text{Var}(x_{is}) = n^{-1} \sum_{i=1}^{p} g_i = \Theta(1) = \Theta(p^{-r_\phi}) \quad (D.11)$$

$$\Longleftrightarrow \tau_b = 0$$
Using this result, we obtain the asymptotic behavior of $A_{kk}$:

$$A_{kk} = \sum_{i=1}^{p} \mathbb{E} \left[ (\hat{\mu}_i^k - \bar{\mu}_i)^2 \right] = \sum_{i=1}^{p} \mathbb{E} \left[ (\hat{\mu}_i^k)^2 - 2\hat{\mu}_i^k \bar{\mu}_i - \bar{\mu}_i^2 \right] \quad (D.12)$$

$$= \sum_{i=1}^{p} \left\{ \mathbb{E} \left[ (\hat{\mu}_i^k)^2 \right] + \mathbb{E} \left[ \bar{\mu}_i^2 \right] \right\} = \sum_{i=1}^{p} \left\{ (\hat{\mu}_i^k)^2 + \text{Var}(\hat{\mu}_i^k) + \text{Var}(\bar{\mu}_i^k) \right\}$$

$$= \Theta(p^{r^k_k}) + \Theta(1) = \Theta(p^{r^k_k})$$

$$\iff \tau^k_k = \max(\tau^k_k, 0)$$

**G3, part I: Consistency of $\hat{A}_{kl}$** As $\hat{A}_{kl}$ is unbiased, I have to show that

$$\text{Var}(\hat{A}_{kl}) = o(p^{r^k_k + r^l_l}) = o(p^{\max(r^k_k, 0) + \max(r^l_l, 0)}) \quad (D.13)$$

I introduce the notation

$$\hat{x}_{is}^{(k)} = x_{is}^{(k)} - \mu_i^{(k)},$$

$$\hat{\mu}_i^{(k)} = n^{-1} \sum_s \hat{x}_{is}^{(k)}.$$

We then have

$$\text{Var}(\hat{A}_{kl}) = \text{Var}(\sum_{i=1}^{p} (\hat{\mu}_i^{k} - \hat{\mu}_i) (\hat{\mu}_i^{l} - \hat{\mu}_i))$$

$$= \text{Var}(\sum_{i=1}^{p} (\hat{\mu}_i^{k} - \bar{\mu}_i + \mu_i^{k}) (\hat{\mu}_i^{l} - \bar{\mu}_i + \mu_i^{l})) \quad (D.14)$$

To show eq. (D.13), it is sufficient to show that the variance of each combination of terms in eq. (D.14) is $o(p^{r^k_k + r^l_l})$. There are three non-constant types of combinations: First, there is the product of a mean and a sample mean:

$$\text{Var} \left( \sum_{i=1}^{p} \mu_i^{k} \hat{\mu}_i^{l} \right) = n_i^{-2} \sum_{ij} \text{Cov}(\mu_i^{k}, \sum_s^l \hat{x}_{is}^{l}, \mu_j^{k}, \sum_t^k \hat{x}_{jt}^{k}) = n_i^{-1} \mu^{kT} C^{l} \mu^{k}$$

$$\overset{(M1)}{=} \frac{\mu^{kT} C^{l} \mu^{k}}{||\mu^{k}||^2} \Theta(p^{r^k_k - 1}) = \max_i \gamma^{l}_i \Theta(p^{r^k_k - 1})$$

$$\overset{(M2), (D2)}{=} o(p^{r^k_k + 1}) \Theta(p^{r^l_l - 1}) = o(p^{r^k_k + r^l_l}) = o(p^{r^k_k + r^l_l})$$

Second, there are products of two different sample means:

$$\text{Var} \left( \sum_{i=1}^{p} \hat{\mu}_i^{k} \hat{\mu}_i^{l} \right) = n^{-2} n_k^{-2} \text{Var} \left( \sum_{i=1}^{p} \sum_s^l \hat{x}_{is}^{l} \hat{x}_{it}^{k} \right) = n^{-1} n_k^{-1} \sum_{i,j=1}^{p} \text{Cov} (x_{i1}, x_{j1}) \text{Cov} (x_{i1}^{k}, x_{j1}^{k})$$

$$= n^{-1} n_k^{-1} \sum_{i,j=1}^{p} \text{Cov} (y_{i1}, y_{j1}) \text{Cov} (z_{i1}^{k}, z_{j1}^{k}) = n^{-1} n_k^{-1} \sum_{i=1}^{p} \gamma_i \mathbb{E}[(\hat{z}_{i1}^{k})^2]$$

$$\leq \frac{1}{nn_k} \sum_{i=1}^{p} \gamma_i \gamma_i^{k} \leq \frac{1}{nn_k} \sqrt{p^{-1} \sum_{i=1}^{p} \gamma_i^2} \sqrt{p^{-1} \sum_{i=1}^{p} (\gamma_i^k)^2}$$
D.3. PROOF OF THEOREM 4.3 (LDL CONSISTENCY OF MTS OF THE MEAN)

equals \( \Theta \left( p^{0.5 (r_n^k + r_n^l)} - 1 \right) \) (M2) \( \Rightarrow o \left( p^{\max(0,r_n^k) + \max(0,r_n^l)} \right) = o(p^{r_n^k + r_n^l}) \)

The third combination has two sample means:

\[
\text{Var} \left( \sum_{i=1}^{p} \tilde{y}_{it}^2 \right) = n^{-4} \text{Var} \left( \sum_{i=1}^{p} \sum_{s,t} y_{is}y_{it} \right) = n^{-4} \sum_{i,j=1}^{p} \sum_{s,t} \text{Cov} (y_{is}y_{it}, y_{js}y_{jt})
\]

\[
= n^{-4} \sum_{i,j=1}^{p} \left\{ \sum_{s} \text{Cov} (y_{is}^2, y_{js}^2) + \sum_{s \neq s'} \text{Cov} (y_{is}y_{it}, y_{js}y_{jt}) \right\}
\]

\[
\leq n^{-3} \sum_{i,j=1}^{p} \text{Cov} (y_{i1}^2, y_{j1}^2) + n^{-2} \sum_{i,j=1}^{p} \text{Cov}^2 (y_{i1}, y_{j1})
\]

\[
\leq p^2 n^{-3} \left( p^{-1} \sum_{i=1}^{p} \sqrt{E [y_{i1}^4]} \right)^2 + pn^{-2} \left( p^{-1} \sum_{i=1}^{p} \gamma_i^2 \right)
\]

\[
= o(p^{r_n^k + r_n^l}) \quad \forall k, l \quad (D.15)
\]

I have shown that the variance of all terms and hence \( \text{Var}(\tilde{A}_{kl}) \) is \( o(p^{r_n^k + r_n^l}) \).

(G3), part II: Consistency of \( \tilde{b} \) The estimator \( \tilde{b} \) is also unbiased, hence I have to show

\[
\text{Var}(\tilde{b}) = o(p^{\gamma}) = o(1).
\]

In a first step, I reformulate the variance:

\[
\text{Var}(\tilde{b}) = \text{Var} \left( \sum_{i=1}^{p} \tilde{\mu}_{i} \right) = \text{Var} \left( n^{-1} (n-1)^{-1} \sum_{i=1}^{n} \sum_{i=1}^{n} (x_{it} - \tilde{\mu}_i)^2 \right)
\]

\[
= n^{-2} (n-1)^{-2} \text{Var} \left( \sum_{i=1}^{p} \sum_{i=1}^{n} x_{it}^2 - n^{-1} \sum_{i=1}^{n} \sum_{i=1}^{n} x_{is}x_{it} \right).
\]

The variance is \( o(1) \) if the variances of both terms in the sum are \( o(p^4) \). I start with

\[
\text{Var} \left( \sum_{i=1}^{p} \sum_{i=1}^{n} x_{it}^2 \right) = n \text{Var} \left( \sum_{i=1}^{p} x_{it}^2 \right) = n \text{Var} \left( \sum_{i=1}^{p} y_{it}^2 \right)
\]

\[
= n \sum_{i,j=1}^{p} \text{Cov} (y_{i1}^2, y_{j1}^2) \leq n \sum_{i,j=1}^{p} \sqrt{E [y_{i1}^4]} \sqrt{E [y_{j1}^4]}
\]

\[
\leq p^2 n (1 + \alpha_4) \left( p^{-1} \sum_{i=1}^{p} \gamma_i^2 \right)^2 \quad \text{(D1)} \quad \Rightarrow O(p^3) = o(p^4).
\]

The variance of the second term in the sum is, following the steps in eq. (D.15),

\[
\text{Var} \left( n^{-1} \sum_{i=1}^{p} \sum_{i=1}^{n} x_{is}x_{it} \right) = O(p^3) + o \left( p^{2 \max(0, \min_k \tau_n^k) + 2} \right) \quad \text{(M1)} \quad \Rightarrow o(p^4).
\]

This concludes the proof the \( \text{Var}(\tilde{b}) = o(p^{\gamma}) = o(1) \).
(G4): Restriction on linear combinations

Let \( L = \mathbb{R}^p \) or \( \mathbb{R}^p \). We have

\[
\Theta \left( p \tau^k_A \right) \overset{!}{=} \min_{\alpha_k \in L} \sum_{i=1}^q E \left[ \left( \sum_{i=1}^K \alpha_l \left( \hat{T}_i - \hat{\theta}_i \right) \right)^2 \right] = \min_{\alpha_k \in L} \sum_{i=1}^q E \left[ \left( \sum_{i=1}^K \alpha_l \left( \hat{\mu}_i - \mu_i \right) \right)^2 \right]
\]

\[
= \min_{\alpha_k \in L} \sum_{i=1}^q \left\{ \sum_{i=1}^K \alpha_l (\hat{\mu}_i - \mu_i)^2 + \sum_{i=1}^K |\alpha_l| \text{Var} (\hat{\mu}_i) + \sum_{i=1}^K |\alpha_l| \text{Var} (\hat{\mu}_i) \right\} \geq \Theta \left( p \tau^k_A \right) + \sum_{i=1}^q \text{Var}(\hat{\mu}_i) = \Theta \left( p \tau^k_A \right) + \Theta(1) = \Theta \left( p \max(0, \tau^k_A) \right)
\]

This concludes the proof of Theorem 4.3.

\[\square\]

D.4 Proof of Theorem 4.4 (FOLDL consistency of MTS of the mean)

Proof. As above, without loss of generality, I assume \( \mu = 0 \). I again start by analyzing the asymptotic behavior of \( \Delta \hat{\theta} \), \( A_{kk} \) and \( b \), then I prove consistency of \( \hat{A}_{kl} \) and \( \hat{b} \).

(G1) & (G2): Asymptotic behavior of \( \Delta \hat{\theta} \), \( A_{kk} \) and \( b \)

From equations (D.11) and (D.12) we directly obtain

\[ \tau^k_\theta = 1 \quad \text{and} \quad \forall k : \tau^k_A = 1. \]

G3, part I: Consistency of \( \hat{A}_{kl} \)

As for the LDL, I show that all types of terms in eq. (D.14) are \( o(p^{\tau^k_A + \tau^k_A}) \). For the FOLDL, this means they have to be \( o(p^2) \). Following similar steps as above, we obtain

\[
\text{Var} \left( \sum_{i=1}^p \hat{\mu}_i \right) = n^{-1} \mu^k \mathbf{C}^k \mu^k = \max_i \gamma_i^k \Theta(p^{\tau^k_A}) \overset{(M2')}{=} o(p^{1 + \tau^k_A}) = o(p^2)
\]

\[
\text{Var} \left( \sum_{i=1}^p \hat{\mu}_i \right) \leq \frac{p}{m^2} \sqrt{p-1} \left( \sum_{i=1}^p \gamma_i^k \right)^2 = o \left( p \max(1, \tau^k_A) + \max(1, \tau^k_A) \right) = o(p^2)
\]

\[
\text{Var} \left( \sum_{i=1}^p \hat{\mu}_i \right) = n^{-4} \sum_{i,j=1}^p \left\{ \sum_s \text{Cov} \left( y_{is}^{y_{js}} y_{js} \right) + \sum_{s,s \neq s} \text{Cov} \left( y_{is} y_{js}, y_{is} y_{js} \right) \right\}
\]

\[
\leq \frac{1}{n^3} \sum_{i,j \neq i} \text{Cov} \left( y_{ik}^2, y_{jk}^2 \right) + \frac{1 + \alpha_4 \alpha_4}{n^3} \left( p^{-1} \sum_{i=1}^{p} \gamma_i^k \right) + \frac{p}{n^2} \left( p^{-1} \sum_{i=1}^{p} \gamma_i^k \right) \overset{(M2'),(M4)}{=} o \left( p^2 + o(p^2) \right)
\]

I have shown that the variance of all terms and hence \( \text{Var}(\hat{A}_{kl}) \) is \( o(p^{\tau^k_A + \tau^k_A}) \).

G3, part II: Consistency of \( \hat{b} \)

I have to show that \( \text{Var}(\hat{b}) = o(p^{2 \tau^k_A}) = o(p^2) \):

\[
\text{Var}(\hat{b}) = \text{Var} \left( \sum_{i=1}^p \text{Var}(\hat{\mu}_i) \right) = \text{Var} \left( \frac{1}{n(n-1)} \sum_{i=1}^n \sum_{t=1}^n (x_{it} - \hat{\mu}_i)^2 \right)
\]
D.5. PROOF OF THEOREM 4.5 (LDL CONSISTENCY OF MTS OF THE COVARIANCE)

\[ \frac{1}{n^2(n-1)^2} \text{Var} \left( \sum_{i=1}^{p} \sum_{t=1}^{n} x_{it}^2 - n^{-1} \sum_{i=1}^{p} \sum_{s,t,s \neq s=1} x_{is}x_{it} \right). \]  \hspace{1cm} (D.18)

This variance expression is \( o(p^2) \) if the variance of each of the three sums is \( o(p^2) \). For the first sum in eq. (D.18), I use eq. (D.17) and obtain

\[ \text{Var} \left( \sum_{i=1}^{p} y_{it}^2 \right) = \sum_{ij} \text{Cov} \left( y_{it}^2, y_{jt}^2 \right) \leq \frac{1}{n^3} \sum_{i,j \neq i} \text{Cov} \left( (y_{ik})^2, (y_{js})^2 \right) + \frac{(1 + \alpha_4)p}{n^3} \left( p^{-1} \sum_{i=1}^{p} \gamma_i^2 \right) \]

\[ = o(p^2) + O(p). \]

The second sum in eq. (D.18) is proportional to the first sum. For the third sum in eq. (D.18), we obtain, by using eq. (D.17),

\[ \text{Var} \left( \sum_{i=1}^{p} \sum_{s,i \neq i} x_{is}x_{it} \right) = \text{Var} \left( \sum_{ij} \sum_{s,t,s \neq s} x_{is}x_{it}x_{js}x_{jt} \right) = o \left( p^2 \right). \]

This concludes the proof that \( \text{Var}(\hat{b}) = o(p^2) = o(p^2) \).

(G4): Restriction on linear combinations

Similar to eq. (D.16), we have

\[ \Theta \left( p^{\tau_{\hat{a}}} \right) = \min_{\alpha \in \Re} \sum_{i=1}^{q} \left( \sum_{l=1}^{K} \alpha_l (\hat{T}_l - \hat{\Theta}_l) \right)^2 \geq \sum_{i} \text{Var}(\hat{\mu}_i^k) = \Theta(p). \]

This concludes the proof of Theorem 4.4. \( \square \)

D.5 Proof of Theorem 4.5 (LDL consistency of MTS of the covariance)

Proof. The estimators \( \hat{A} \) and \( b \) depend on the choice of target. I restrict the proof on targets given by sample covariance matrices of additional data sets. The biased estimators in Schäfer & Strimmer (2005) and Ledoit & Wolf (2003b) have smaller variance, consistency can be shown following similar steps.

(G1) & (G2): Asymptotic behavior of \( \Delta^{\hat{a}}, b_k \) and \( A_{kk} \)

I first show the asymptotic behavior

\[ \Delta^{\hat{a}} = b_k = \left( \sum_{ij} \text{Var}(S_{ij}) \right) = \Theta(p) \overset{\dagger}{=} \Theta(p^{\tau_{\hat{a}}}) \iff \tau_{\hat{a}} = 1. \]  \hspace{1cm} (D.19)

Rotation invariance allows us to analyze in the eigenbasis. The upper bound follows from

\[ \sum_{i,j} \text{Var}(S_{ij}) \leq \frac{1}{n} \sum_{i,j} \left\{ \sqrt{\text{Var}(y_{i1}^2) \text{Var}(y_{j1}^2)} \right\} \]

\[ + \text{E} \left[ y_{i1}^2 \right] \text{E} \left[ y_{j1}^2 \right] - \text{E}^2 \left[ y_{i1}y_{j1} \right] \]  \hspace{1cm} (D.20)

\[ \leq \frac{1}{n} \sum_{i,j} \left\{ \sqrt{\text{E}[y_{i1}^4] \text{E}[y_{j1}^4]} + \text{E} \left[ y_{i1}^2 \right] \text{E} \left[ y_{j1}^2 \right] - \text{E}^2 \left[ y_{i1}y_{j1} \right] \right\} \]
\[ \leq \frac{2}{n} \sum_{i,j} \sqrt{\mathbb{E}[y_{i1}^4]\mathbb{E}[y_{j1}^4]} \leq \frac{2p^2}{n} (1 + \alpha_4) \left( \frac{1}{p} \sum_{i} \mathbb{E}[y_{i1}^2] \right)^2 = \Theta(p). \]

For the lower bound, I distinguish two cases: for \( \tau_\gamma = 1 \), we have
\[
\sum_{i,j} \text{Var}(S'_{ij}) \geq \frac{1}{n} \sum_{i,j} \{ \mathbb{E} [y_{i1}^4] - \mathbb{E}^2 [y_{i1}^2] \} \geq \frac{1}{n} \sum_{i,j} \{ \mathbb{E} [y_{i1}^2] \mathbb{E} [y_{j1}^2] - \mathbb{E}^2 [y_{i1} y_{j1}] \} = \Theta(p) - \Theta(p^\gamma) = \Theta(p).
\]

For the case \( \tau_\gamma < 1 \), we have
\[
\sum_{i,j} \text{Var}(S'_{ij}) = \frac{1}{n} \sum_{i,j} \{ \mathbb{E} [y_{i1}^2 y_{j1}^2] - \mathbb{E}^2 [y_{i1} y_{j1}] \} \geq \frac{1}{n} \sum_{i,j} \{ \mathbb{E} [y_{i1}^4] \mathbb{E} [y_{j1}^2] - \mathbb{E}^2 [y_{i1} y_{j1}] \} = \Theta(p) - \Theta(p^\gamma) = \Theta(p).
\]

The asymptotic behavior of \( A_{kk} \) depends on the relationship between the original data \( X \) and the additional data set \( X^k \):
\[
A_{kk} = \sum_{i,j=1}^{p} \mathbb{E} [(S_{ij}^k - S_{ij}) (S_{ij}^k - S_{ij})],
\]
\[
= \sum_{i,j=1}^{p} \mathbb{E} [(C_{ij} - C_{ij}^k) S_{ij}^k],
\]
\[
= \mathbb{E} [C_{ij} S_{ij}^k] + \text{Var}(S_{ij}) \overset{(G1)}{=} \mathbb{E} [p^\gamma] + \Theta(p^\gamma) = \Theta(p^\gamma)
\]
\[
\iff \tau_A^k = \max(1, \tau_\gamma^k).
\]

**G3: Consistency of \( \hat{A}_{kl} \)** As the estimator \( \hat{A}_{kl} \) is unbiased (Bart \& Müller, 2013), I have to show that
\[
\text{Var} \left( \hat{A}_{kl} \right) = \text{Var} \left( \sum_{i,j=1}^{p} (S_{ij}^k - S_{ij}) (S_{ij}^k - S_{ij}) \right)
\]
\[
= \text{Var} \left( \sum_{i,j=1}^{p} S_{ij} S_{ij}^k - S_{ij}^k S_{ij}^k - S_{ij} S_{ij}^k + S_{ij}^2 \right),
\]
\[
= o(p_{\max}^\gamma + \tau_A^k) = o(p_{\max}^\gamma + \tau_\gamma^k).
\]

It suffices to show that the variances of all terms in the sum in eq. (D.23) are \( o(p_{\max}^\gamma + \tau_\gamma^k) \).

**Variance of \( \sum_{ij} S_{ij}^2 \)** I start with the product of two identical sample covariances:
\[
\sum_{ij} S_{ij}^2 = \sum_{ij} \left( \frac{1}{n} \sum_{s} y_{is} y_{js} \right) = \frac{p^2}{n^2} \sum_{ts} \left( \frac{1}{p} \sum_{i} y_{is} y_{it} \right)^2
\]
\[
= \frac{p^2}{n^2} \sum_{s} \left( \frac{1}{p} \sum_{i} y_{is}^2 \right)^2 + \frac{1}{n^2} \sum_{s,t \neq s} \left( \sum_{i} y_{is} y_{it} \right)^2.
\]
Again, it is sufficient to show that the variance of both terms separately is \(o(p^{r+\tau})\). For the first term, we have

\[
\text{Var} \left( \frac{p^2}{n^2} \sum_i \left( \frac{1}{p} \sum_i y_{is}^2 \right)^2 \right) \leq \frac{p^4}{n^3} \mathbb{E} \left[ \left( \frac{1}{p} \sum_i y_{i1}^2 \right)^4 \right] \leq \frac{p^4(1 + \alpha)}{n^3} \mathbb{E} \left[ \frac{1}{p} \sum_i y_{i1}^2 \right] = \mathcal{O}(p) = o(p^{r+\tau}).
\]

Let us now look at the second term in eq. (D.24):

\[
\text{Var} \left( \frac{1}{n^2} \sum_{s,t \neq s} \left( \sum_i y_{is} y_{it} \right)^2 \right) = \frac{1}{n^2} \sum_{s,t \neq s} \sum_{s',t' \neq s'} \text{Cov} \left( \left( \sum_i y_{is} y_{it} \right)^2, \left( \sum_i y_{is'} y_{it'} \right)^2 \right). \tag{D.25}
\]

The covariance expression only depends on the cardinal of the intersection, which I denote by \((\{s, t\} \cup \{s', t'\})^\#\) and which can take the values of 0, 1 and 2.

**Case 1: Cardinality equal to zero.** When this cardinality is zero,

\[
(\{s, t\} \cup \{s', t'\})^\# = 0,
\]

there is independence and the covariance is zero as well.

**Case 2: Cardinality equal to one.** For

\[
(\{s, t\} \cup \{s', t'\})^\# = 1,
\]

we have \(\Theta(n^3)\) expressions of the form

\[
\text{Cov} \left( \left( \sum_i y_{i1} y_{i2} \right)^2, \left( \sum_i y_{i1} y_{i3} \right)^2 \right) = \mathbb{E} \left[ \left( \sum_i y_{i1} y_{i2} \right)^2 \left( \sum_i y_{i1} y_{i3} \right)^2 \right] - \mathbb{E} \left[ \left( \sum_i y_{i1} y_{i2} \right)^2 \right] \mathbb{E} \left[ \left( \sum_i y_{i1} y_{i3} \right)^2 \right] \leq \max \left( \mathbb{E} \left[ \left( \sum_i y_{i1} y_{i2} \right)^2 \left( \sum_i y_{i1} y_{i3} \right)^2 \right], \mathbb{E}^2 \left[ \left( \sum_i y_{i1} y_{i2} \right)^2 \right] \right),
\]

as both terms are positive. For the first term, we have

\[
\mathbb{E} \left[ \left( \sum_i y_{i1} y_{i2} \right)^2 \left( \sum_i y_{i1} y_{i3} \right)^2 \right] = \sum_{i,j,i',j'} \mathbb{E} [y_{i1} y_{i1} y_{j1} y_{j1} y_{i1} y_{i1} y_{j1} y_{j3} y_{j3} y_{j3}] \mathbb{E} [y_{i2} y_{i2} y_{j2}] \mathbb{E} [y_{j3} y_{j3}]
\]

\[
= \sum_{i,j} \mathbb{E} [y_{i1}^2 y_{j1}^2] \mathbb{E} [y_{i2}^2] \mathbb{E} [y_{j3}^2] \leq p^2 \left( \frac{1}{p} \sum_i \mathbb{E} [y_{i1}^4] \mathbb{E} [y_{i2}^2] \right)^2 \leq p^2 (1 + \alpha_4) \left( \frac{1}{p} \sum_i \mathbb{E} [y_{i1}^2] \right)^2 = \mathcal{O} \left( p^{2\tau+2} \right). \tag{D.26}
\]

For the second term, we have

\[
\mathbb{E}^2 \left[ \left( \sum_i y_{i1} y_{i2} \right)^2 \right] = \left( \sum_{i,j} \mathbb{E} [y_{i1} y_{j1}] \right)^2 \leq p^2 \left( \frac{1}{p} \sum_i \mathbb{E} [y_{i1}^2] \right)^2 = \mathcal{O} \left( p^{2\tau+2} \right) \tag{D.27}
\]
The results eq. (D.26) and eq. (D.27) yield, plugged into eq. (D.25),

\[
\frac{\Theta(n^3)}{n^4} \left| \mathrm{Cov} \left( \left( \frac{1}{p} \sum_i y_{i1} y_{i2} \right)^2, \left( \frac{1}{p} \sum_i y_{i3} y_{i4} \right)^2 \right) \right| = \frac{1}{n} \mathcal{O} \left( p^{2\tau_{+} + 2} \right) = O \left( p^{2\tau_{+} + 1} \right) \overset{(C3)}{=} o(p^{\tau_{A} + \tau_{A}'}),
\]

therefore I have shown that the terms with \( \{\{s, t\} \cup \{s', t'\}\}^{\#} = 1 \) are \( o(p^{\tau_{A} + \tau_{A}'}) \).

**Case 3: Cardinality equal to two.** For

\( \{\{s, t\} \cup \{s', t'\}\}^{\#} = 2 \),

we get \( \Theta(n^2) \) expressions of the form

\[
\sum_{i,j,i',j'} |\mathrm{Cov} (y_{i1} y_{i2} y_{i'1} y_{i'2}, y_{j1} y_{j2} y_{j'1} y_{j'2})|.
\]

I decompose the set of integers into two disjoint subsets: \( \{1, \ldots, p\}^4 = Q \cup R \), where \( Q \) is the set of distinct integers and \( R \) is the remainder:

\[
= \sum_{i,j,i',j' \in Q} |\mathrm{Cov} (y_{i1} y_{i2} y_{i'1} y_{i'2}, y_{j1} y_{j2} y_{j'1} y_{j'2})| + \sum_{i,j,i',j' \in R} |\mathrm{Cov} (y_{i1} y_{i2} y_{i'1} y_{i'2}, y_{j1} y_{j2} y_{j'1} y_{j'2})|.
\]

The sum over \( Q \) I can bring into a form which is dominated as a consequence of (C2):

\[
|\mathrm{Cov} (y_{i1} y_{i2} y_{i'1} y_{i'2}, y_{j1} y_{j2} y_{j'1} y_{j'2})| = |E^2 [y_{i1} y_{i'1} y_{j1} y_{j'1}] - E^2 [y_{i1} y_{i'1}] E^2 [y_{j1} y_{j'1}]| = E^2 [y_{i1} y_{i'1} y_{j1} y_{j'1}]
\]

\[
= (\mathrm{Cov} (y_{i1} y_{i'1}, y_{j1} y_{j'1}) + E [y_{i1} y_{i'1}] E [y_{j1} y_{j'1}])^2 = (\mathrm{Cov} (y_{i1} y_{i'1}, y_{j1} y_{j'1}))^2.
\]

Taking the pre-factors into account, we get

\[
\frac{\Theta(n^2)}{n^4} \sum_{(i,j,i',j') \in Q} |\mathrm{Cov} (y_{i1} y_{i2} y_{i'1} y_{i'2}, y_{j1} y_{j2} y_{j'1} y_{j'2})| \leq \Theta(p^2) \frac{ \sum_{(i,j,i',j') \in Q} (\mathrm{Cov} (y_{i1} y_{i'1}, y_{j1} y_{j'1}))^2 }{|Q_p|} \overset{(C2)}{=} o(p^2) = o(p^{\tau_{A} + \tau_{A}'}).
\]

For the sum over \( R \), we have

\[
\sum_{(i,j,i',j') \in R} |\mathrm{Cov} (y_{i1} y_{i2} y_{i'1} y_{i'2}, y_{j1} y_{j2} y_{j'1} y_{j'2})|
\]

\[
\leq \sum_{i,j,i',j'} 2 |\mathrm{Cov} (y_{i1}^2 y_{i2}^2, y_{j1} y_{j2} y_{j'1} y_{j'2})| + 4 |\mathrm{Cov} (y_{i1} y_{i2} y_{i'1} y_{i'2}, y_{j1} y_{j2} y_{j1} y_{j2})|
\]

\[
\leq \sum_{i,j,i',j'} 2 \sqrt{E [y_{i1}^4] E [y_{i2}^2 y_{j1}^2 y_{j2}^2]} + 4 \sqrt{E [y_{i1} y_{i2}^2 y_{i'1}^2 y_{i'2}^2] E [y_{j1}^2 y_{j2} y_{j'1}^2 y_{j'2}^2]}
\]

\[
\leq 6 \sum_{i,j,i',j'} E [y_{i1}^4] \sqrt{E [y_{i1}^4]} \leq 6 p^3 (1 + \alpha_4) \left( \frac{1}{p} \sum_i E^2 [y_{i1}^2] \right) \left( \frac{1}{p} \sum_j E [y_{j1}^2] \right)^2
\]

\[
= O (p^{2\tau_{+} + 3}).
\]

(D.29)
Together with the pre-factors, we obtain
\[ \frac{2n(n - 1)}{n^4} \sum_{(i, j, s, t) \in R} |\text{Cov}(y_{11}y_{12}y_{13}y_{14}, y_{21}y_{22}y_{23}y_{24})| = \frac{1}{n^2} \mathcal{O}(p^{2\gamma_1 + 3}) = \mathcal{O}(p^{2\gamma_1 + 1}) \quad \text{(C3)} = o(p^{\tau_k + \tau_h}). \]

This finishes the proof for the terms with \((s, t) \cup \{s', t'\} \neq 2\) and in total I have shown that \(\text{Var}(\sum_{ij} S_{ij}^2)\) is \(o(p^{\tau_k + \tau_h})\). For \(\text{Var}(S_{ij}^k S_{ij}^k), k = l\), an analogue proof holds.

**Variance of \(\sum_{ij} S_{ij}^k S_{ij}\)** Let us now analyze the products of different sample covariances in eq. (D.23).

\[
\text{Var} \left( \sum_{ij} S_{ij}^k S_{ij} \right) = \text{Var} \left( \sum_{ij} \sum_{st} x_{it}^k x_{jt}^k \right) = \frac{1}{n^2} \sum_{ijgh} \text{Cov} \left( x_{ij}^k x_{ij}^k, x_{ij}^k x_{ij}^k \right) = \frac{1}{n^2} \sum_{ijgh} \text{Cov} \left( x_{ij}^k x_{ij}^k, x_{ij}^k x_{ij}^k \right) - C_{ij}^k C_{gh}^k \text{Cov} \left( x_{ij}^k x_{ij}^k, x_{ij}^k x_{ij}^k \right).
\]

The first term can be separated into the contributions from the two different data sets:

\[
\frac{1}{n^2} \sum_{ijgh} \text{Cov} \left( x_{ij}^k x_{ij}^k, x_{ij}^k x_{ij}^k \right) \text{Cov} \left( x_{ij}^k x_{ij}^k, x_{ij}^k x_{ij}^k \right) \leq \frac{1}{n^2} \sum_{ijgh} \text{Cov}^2 \left( x_{ij}^k x_{ij}^k, x_{ij}^k x_{ij}^k \right) + \text{Cov}^2 \left( x_{ij}^k x_{ij}^k, x_{ij}^k x_{ij}^k \right).
\]

These terms are rotation invariant, therefore we analyze

\[
\frac{1}{n^2} \sum_{ijgh} \text{Cov}^2 \left( y_{ij}, y_{ij} \right).
\]

For \(i, j, g, h\) distinct, this leads directly to assumption (C2). Otherwise, we have

\[
\frac{1}{n^2} \sum_{ijgh} \text{Cov}^2 \left( y_{ij}, y_{ij} \right) \leq \frac{4}{n^2} \sum_{ij} \text{Cov}^2 \left( y_{ij}, y_{ij} \right) + \text{Cov}^2 \left( y_{ij}, y_{ij} \right) \\
\leq \frac{8}{n} \sum_{ij} \mathbb{E} \left[ (y_{ij})^2 \right] \mathbb{E} \left[ (y_{ij})^4 \right] \\
\leq \frac{8}{n} \sum_{ij} \sqrt{\mathbb{E} \left[ (y_{ij})^4 \right]} \sqrt{\mathbb{E} \left[ (y_{ij})^4 \right]} \\
\leq \frac{8 \rho_3}{n} \left( \frac{1}{p} \sum_i \gamma_i \right)^2 \left( \frac{1}{p} \sum_g \gamma_i \right) \\
= \mathcal{O}(p^{\tau_1 + 1}) = o(p^{\tau_1 + \tau_h}).
\]

Next we consider the second term,

\[
\frac{1}{n^2} \sum_{ijgh} C_{ij}^k C_{gh}^k \text{Cov} \left( x_{ij}^k x_{ij}^k, x_{ij}^k x_{ij}^k \right) = \frac{1}{n^2} \sum_{ijgh} \Sigma_{ij}^k \Sigma_{gh}^k \text{Cov} \left( z_{ij}, z_{ij} \right) \\
\leq \frac{1}{n^2} \sum_{ij} \gamma_i \gamma_g \sqrt{\mathbb{E} \left[ z_i^4 \right]} \sqrt{\mathbb{E} \left[ z_g^4 \right]} = \frac{1}{n^2} \left( \sum \gamma_i \sqrt{\mathbb{E} \left[ z_i^4 \right]} \right) \\
\leq \frac{p^3}{n^2} \left( \frac{1}{p} \sum_i \gamma_i^2 + \mathbb{E} \left[ z_i^4 \right] \right)^2 \leq \frac{p^3(1 + \alpha_2)}{n^2} \left( \frac{1}{p} \sum_i \gamma_i^2 + \gamma_i^2 \right)^2
\]
\[ \text{APPENDIX D. PROOFS FOR MULTI-TARGET SHRINKAGE} \]

\[ \mathcal{O}(p^{2 \max(\tau, \tau_k^*)}) = o(p^{r_k^* + \tau_k^*}). \]

With this I have shown that all terms in \( \text{Var} \left( \sum_{ij} S_{ij}^k \right) \) and hence \( \text{Var}(\hat{A}_{kl}) \) is \( o(p^{r_k^* + \tau_k^*}). \)

(G3), part II: Consistency of \( \hat{b}_k \) By reformulation we obtain

\[
\sum_{ij} \text{Var} S_{ij} = \sum_{ij} \left( \frac{1}{(n-1)n} \sum_s \left( y_{is}y_{js} - \frac{1}{n} \sum_{s'} y_{is'y_{js'}} \right)^2 \right) = \frac{1 - \frac{1}{n}}{(n-1)n} \sum_{ij} \left( \sum_s y_{is}^2 y_{js}^2 - \frac{1}{n} \sum_{ss'} y_{is} y_{is'} y_{js} y_{js'} \right)
= \frac{p^2}{(n-1)n} \sum_s \left( \frac{1}{p} \sum_i y_{is}^2 \right)^2 - \frac{1}{(n-1)} \sum_{ij} S_{ij}^2. \tag{D.30}
\]

Both terms, with different pre-factors, have been analyzed above. The variance of first term is \( \mathcal{O}(p) \) and the bound on the variance of the second term is \( n^{-2} o(p^{2 \max(1, \tau_c)}) = o(p^2). \) Hence \( \text{Var}(\hat{b}) \) is \( o(p^{2 \tau_k}). \)

(G4): Restriction on linear combinations Let \( L \) be \( \mathbb{R}^p \) or \( \mathbb{R}_{\geq 0}^p. \) We have

\[
\Theta \left( p^{r_k^*} \right) = \min_{\alpha \in L} \sum_i \mathbb{E} \left[ \left( \sum_{i=1}^K \alpha_i (T_i^l - \hat{\theta}_i) \right)^2 \right] = \min_{\alpha \in L} \sum_i \mathbb{E} \left[ \left( \sum_{i=1}^K \alpha_i (S_i^l - S_i) \right)^2 \right] \tag{D.31}
= \min_{\alpha \in L} \sum_i \left( \sum_{i=1}^K \alpha_i (S_i^l - S_i) \right)^2 + \text{Var} \left( \sum_{i=1}^K \alpha_i S_i \right) + \text{Var} \left( \sum_{i=1}^K \alpha_i S_i \right) \geq \Theta \left( p^{r_k^*} \right) + \sum_i \text{Var}(S_i) = \Theta \left( p^{r_k^*} \right) + \Theta(p) = \Theta \left( p^{\max(1, \tau_k^*)} \right)
\]

This concludes the proof of Theorem 4.5. \( \square \)

D.6 Proof of Theorem 4.6 (FOLDL consistency of MTS of the covariance)

(G1) & (G2): Asymptotic behavior of \( \Delta^\hat{\theta}, b_k \) and \( A_{kk} \)

Proof. I first show the asymptotic behaviour

\[
\Delta^\hat{\theta} = b_k = \left( \sum_{ij} \text{Var}(S_{ij}) \right) = \Theta(p^2) \overset{\dagger}{=} \Theta(p^{r_k^*}) \iff \tau_\theta = 2 \tag{D.32}
\]

The upper bound follows from (compare to eq. (D.20))

\[
b_k = \sum_{i,j} \text{Var}(S_{ij}) \leq \frac{2p^2}{n} (1 + \alpha_4) \left( \frac{1}{p} \sum_i \mathbb{E}(y_{1i}^2) \right)^2 \leq \Theta(p^2). \]
For the lower bound, I again distinguish two cases: for $\tau = 1$, we have (compare to eq. (D.21))
\[
\sum_{i,j} \text{Var}(S'_{ij}) = \frac{\beta_p}{n} \frac{1}{p} \sum_i \gamma_i^2 = \Theta(p^{1+\tau}) = \Theta(p^2).
\]
For the case $\tau < 1$, we have (compare to eq. (D.22))
\[
\sum_{i,j} \text{Var}(S'_{ij}) \geq \frac{p^2}{n} \left( \frac{1}{p} \sum_i \mathbb{E}[x_{i1}^2] \right)^2 - \frac{p}{n} \frac{1}{p} \sum_i \gamma_i^2 = \Theta(p^2) - \Theta(p^{1+\tau}) = \Theta(p^2).
\]
For the asymptotic behavior of $A_{kk}$ we then have
\[
A_{kk} = \sum_{i,j=1}^p (C_{ij} - C_{ik}^k)^2 + \text{Var}(S'_{ij}) + \text{Var}(S_{ij}) = \Theta(p^{1+\tau}) + \Theta(p^2) = \Theta(p^{1+\tau}), \quad (D.33)
\]
\[\iff \forall k: \tau_A^k = 2\]
where used the fact that $\sum_{ij} \text{Var}(S_{ij}^k)$ has the same limit behavior as $\sum_{ij} \text{Var}(S_{ij})$.

(G3), part I: Consistency of $\hat{A}_{kl}$ The proof is analogue to the proof in Theorem 4.5. I only show that $\text{Var}\left(\sum_{ij} S_{ij}^2\right)$, the expression with the highest variance, is $o(p^{1+\tau_A}) = o(p^4)$. I use the same decomposition as above:
\[
\sum_{ij} S_{ij}^2 = \frac{1}{n^2} \sum_s \left( \sum_i y_{is}^2 \right)^2 + \frac{1}{n^2} \sum_{s,t \neq s} \left( \sum_i y_{is} y_{it} \right)^2. \quad (D.34)
\]
This asymptotic setting is easier, because the sums over $s$ and $t$ are finite sums. We have a finite number of terms in the first sum in eq. (D.34):
\[
\text{Var}\left(\left(\sum_i y_{is}\right)^2\right) = \sum_{i,j',j''} \text{Cov}\left(y_{11}^2 y_{j1}^2, y_{11}^2 y_{j'1}^2\right) = \sum_{i,j',j'' \in Q} \text{Cov}\left(y_{11}^2 y_{j1}^2, y_{11}^2 y_{j'1}^2\right) + \sum_{i,j',j'' \in R} \text{Cov}\left(y_{11}^2 y_{j1}^2, y_{11}^2 y_{j'1}^2\right). \quad (D.35)
\]
For the sum over $Q$, we need assumption (C4):
\[
\sum_{i,j',j'' \in Q} \text{Cov}\left(y_{11}^2 y_{j1}^2, y_{11}^2 y_{j'1}^2\right) \leq p^4 2^4 \sum_{i,j',j'' \in Q} \text{Cov}\left(y_{11}^2 y_{j1}^2, y_{11}^2 y_{j'1}^2\right) = o(p^4). \quad (C4)
\]
For the sum over $R$, we have,
\[
\sum_{(i',j',j'') \in R} \text{Cov}\left(y_{11}^2 y_{j1}^2, y_{11}^2 y_{j'1}^2\right) \leq 6 \sum_{i,j,i'} \text{Cov}\left(y_{11}^2 y_{j1}^2, y_{i1}^2 y_{i'1}^2\right) + \text{Cov}\left(y_{11}^2 y_{j1}^2, y_{i1}^2\right) + \text{Cov}\left(y_{11}^2 y_{j'1}^2, y_{i'1}^2\right)
\]
\[
\leq 6 \sum_{i,j,i'} \sqrt{\mathbb{E}[y_{11}^2] \mathbb{E}[y_{j1}^2]} \sqrt{\mathbb{E}[y_{i1}^2] \mathbb{E}[y_{i'1}^2]} + \sqrt{\mathbb{E}[y_{11}^2] \mathbb{E}[y_{j1}^2]} \sqrt{\mathbb{E}[y_{i'1}^2]}
\]
\[
\leq 6 \sum_{i,j,i'} \sqrt{\mathbb{E}[y_{11}^2] \mathbb{E}[y_{j1}^2]} \sqrt{\mathbb{E}[y_{i1}^2] \mathbb{E}[y_{i'1}^2]} + \sqrt{\mathbb{E}[y_{11}^2] \mathbb{E}[y_{j1}^2]} \sqrt{\mathbb{E}[y_{i'1}^2]}.
\[ \leq 12(1 + \alpha_8) \sum_{i,j,i',j'} \mathbb{E}[y_{i1}^2] \mathbb{E}[y_{j1}^2] \mathbb{E}[y_{i'1}^2] = O(p^{3+\gamma}) \] (C3') \[ = o(p^4). \]

For the terms in the second sum in eq. (D.34), we have

\[
\text{Var} \left( \left( \sum_i y_{i1}y_{i2} \right)^2 \right) = \sum_{i,j,i',j'} \text{Cov} \left( y_{i1}y_{j1}y_{j2}, y_{i'1}y_{i'2}2y_{j'1}y_{j'2} \right)
\]
\[
\leq \sum_{i,j,i',j'} |\text{Cov} \left( y_{i1}y_{j1}y_{i'2}y_{j'1}y_{j'2}, y_{j1}y_{j2}y_{j'1}y_{j'2} \right)|.
\]

For the sum over \(Q\), I simplify using eq. (D.28) and obtain

\[
\sum_{i,j,i',j' \in Q} |\text{Cov} \left( y_{i1}y_{i'1}y_{i'2}, y_{j1}y_{j1}y_{j'1}y_{j'2} \right)| = \sum_{i,j,i',j' \in Q} (\text{Cov} \left( y_{i1}y_{i'1}, y_{j1}y_{j'1} \right))^2 \] (C4)
\[
\leq 24p^4 \sum_{(i,j,i',j') \in Q} |\text{Cov} \left( y_{i1}y_{i'1}, y_{j1}y_{j'1} \right)|^2 \] (C4)
\[
\leq o(p^4).
\]

For the sum over \(R\), we have, as in eq. (D.29),

\[
\sum_{i,j,i',j' \in R} |\text{Cov} \left( y_{i1}y_{i'1}y_{i'2}, y_{j1}y_{j2}y_{j'1}y_{j'2} \right)| = \Theta \left( p^{3+\gamma} \right) \] (C3')
\[
\leq o(p^4).
\]

With this I have shown that all terms and hence Var\((\hat{A}_{kl})\) is \(o(p^{4+k}\tau_\lambda)\).

**G3, part II: Consistency of \(\hat{b}_k\)** As in eq. (D.30) we have

\[
\sum_{ij} \text{Var}(S_{ij}) = \frac{p^2}{(n-1)n} \sum_s \left( \frac{1}{p} \sum_i y_{is} \right)^2 - \frac{1}{(n-1)} \sum_{ij} S_{ij}^2.
\]

The first term is equal to the first term in eq. (D.34) and hence its variance \(o(p^4)\). The second term is proportional to the left hand side of eq. (D.34) and its variance therefore also \(o(p^4)\). In total, Var\((\hat{b})\) is \(o(p^{2+\gamma})\).

**(G4): Restriction on linear combinations** Following the same steps as in eq. (D.31), we obtain

\[
\Theta \left( p^{rA} \right) = \min_{\alpha_k \in R} \sum_{i=1}^q \mathbb{E} \left[ \left( \sum_{l=1}^K \alpha_l (\tilde{T}_i^l - \tilde{\theta}_i) \right)^2 \right] \geq \sum_i \text{Var}(S_i^k) = \Theta(p^2)
\]

This concludes the proof of Theorem 4.6.
Bibliography


Schreuder, M., Rost, T., & Tangermann, M. (2011). Listen, you are writing! Speeding up online spelling with a dynamic auditory BCI. Frontiers in Neuroscience, 5(112).


