Models with Low-Rank and Group-Sparse Components and their Recovery via Convex Optimization

FRIEDRICH-SCHILLER-
UNIVERSITÄT
JENA

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

vorgelegt dem Rat der Fakultät für Mathematik und Informatik
der Friedrich-Schiller-Universität Jena

eingereicht von Frank Nussbaum (M. Sc. Mathematik)

Betreuer: Prof. Dr. Joachim Giesen

Jena, 2021

Gutachter der Dissertation:

1. Professor Dr. Joachim Giesen, FSU Jena
2. Professor Dr. Kristian Kersting, TU Darmstadt
3. Professor Dr. Christopher Schneider, EAH Jena

Tag der öffentlichen Verteidigung: 19.7.2021

Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass

- mir die Promotionsordnung der Fakultät bekannt ist,
- ich die Dissertation selbst angefertigt habe, keine Textabschnitte oder Ergebnisse eines Dritten oder eigenen Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel, persönlichen Mitteilungen und Quellen in meiner Arbeit angegeben habe,
- ich die Hilfe eines Promotionsberaters nicht in Anspruch genommen habe und dass Dritte weder unmittelbar noch mittelbar geldwerte Leistungen von mir für Arbeiten erhalten haben, die im Zusammenhang mit dem Inhalt der vorgelegten Dissertation stehen,
- ich die Dissertation weder anderwaltig als Dissertation bei einer anderen Hochschule noch als Prüfungsarbeit für eine staatliche oder andere wissenschaftliche Prüfung eingereicht habe.

Ort, Datum

Frank Nussbaum

Abstract

In this dissertation, we consider models with low-rank and group-sparse components. First, we investigate robust principal component analysis, where the low-rank component represents the principal components, and the group-sparse component accounts for corruptions in the data. We propose a model for the general setting, where groups of observed variables can be corrupted. Second, we generalize fused latent and graphical models to the class of conditional Gaussian distributions with mixed observed discrete and quantitative variables. Fused latent and graphical models are characterized by a decomposition of the pairwise interaction parameter matrix into a group-sparse component of direct interactions and a low-rank component of indirect interactions due to a small number of quantitative latent variables. All models in this thesis can be learned by solving convex optimization problems with low-rank and group-sparsity inducing regularization terms. For fused latent and graphical models, there is an additional likelihood term. We show that under identifiability assumptions, a given true model can be recovered exactly (principal component analysis) or consistently (fused latent and graphical models, high-dimensional setting) by solving the respective optimization problems. We also present heuristics for selecting the regularization parameters that appear in the optimization problems. We conduct experiments on synthetic and real-world data to support our theory.

Zusammenfassung

Gegenstand dieser Dissertation sind Modelle mit *low-rank* und *group-sparse* Komponenten. Zuerst betrachten wir robuste Hauptachsenanalyse, wobei die low-rank Komponente die Hauptachsen repräsentiert und die group-sparse Komponente Verunreinigungen in den Daten widerspiegelt. Wir untersuchen ein allgemeines Modell, in dem Gruppen von beobachteten Variablen verunreinigt sein können. Als nächstes verallgemeinern wir fusionierte latente und graphische Modelle auf Verteilungen mit sowohl diskreten als auch quantitativen beobachteten Variablen. Fusionierte latente und graphische Modelle sind durch eine Zerlegung der paarweisen Interaktionsparametermatrix in eine group-sparse Komponente für direkte Interaktionen und eine low-rank Komponente für indirekte Interaktionen charakterisiert. Die indirekten Interaktionen hängen mit einer kleinen Zahl an latenten quantitativen Variablen zusammen. Alle Modelle in dieser Dissertation können durch das Lösen von konvexen Optimierungsproblemen mit low-rank und group-sparsity induzierenden Regularisierungstermen gelernt werden. Für fusionierte latente und graphische Modelle kommt ein zusätzlicher Likelihood-Term hinzu. Wir zeigen, dass unter Identifizierbarkeitsannahmen ein wahres Modell durch das Lösen der entsprechenden Optimierungsprobleme exakt (Hauptachsenanalyse) oder konsistent (fusionierte latente und graphische Modelle, im hochdimensionalen Setting) rekonstruiert werden kann. Außerdem stellen wir Heuristiken zur Wahl der in den Optimierungsproblemen auftretenden Regularisierungsparameter vor. Experimente mit synthetischen und realen Daten bestätigen unsere theoretischen Erkenntnisse.

Acknowledgements

First and foremost, I would like to thank my advisor Professor Joachim Giesen. His enthusiasm, wealth of ideas, and clarity of thought have made it an exceptional experience to be working with him. I would like to thank him for the countless conversations about research and life, and for always taking serious interest in my personal development. I consider myself lucky to have had him as an advisor.

I thank Professor Christopher Schneider for being on my reading and defense committees and for providing much valuable feedback. I would also like to thank Professor Kristian Kersting for being on my reading and defense committees, Professor Stefan Ankirchner for acting as the chair of my defense committee, and Professor Martin Mundhenk, Professor Kai Lavonn, and Dr. habil. Sören Laue for being on my defense committee.

It has been a pleasure to have had such wonderful colleagues, both at the FSU (Friedrich-Schiller University) and the DLR (Deutsches Zentrum für Luft- und Raumfahrt). In particular, I thank Andreas, Anna, Auliya, Christoph, Christopher, Jakob, Jonas, Julia, Julien, Lars, Mark, Matthias, Paul, Philipp, and Sören. I have learned a lot from our vivid exchange of ideas, which was also driven by the many individual interests and talents. I made many friends and enjoyed the open, friendly, and enthusiastic work environment. Together, the journey was much more fun.

My sincere acknowledgments go to all teachers and mentors that I had during the various stages of my education. Special thanks goes to my former math teacher Dr. Bernd Licht who encouraged me early and reinforced my decision to study Mathematics. Special thanks also goes to Professor Daniel Lenz from my undergraduate studies. I greatly admire his teaching and scientific skills.

I would like to express my deepest gratitude to my wife Tine. She has been with me during my whole PhD, always ready to support me whenever needed. In these special times, we have not only been sharing a home, but also an office space. Working next to each other, it often turned out that we can talk about aspects from our work in an understanding manner, which even led to new insights. Most importantly however, I would like to thank Tine for her warmth and patience.

I am grateful to all the friends that supported me during the years. In particular, I would like to thank Ben, Jan, Jana, Lisa, Matthias, Max, Richi, Rici, Sebastian, Siggi, Susan, Till, Tobias, and Tonia.

Finally, I would like to thank my family for their unconditional love and support. Especially my parents ensured that me and my two brothers always had access to the best possible education. Mum and dad, I dedicate this thesis to you.

Funding: My position was funded in part by the SDW (German Foundation of the Industries) and by the DLR (general agreement 78 002 077). Moreover, part of the work was done in the context of the German Science Foundation (DFG) grant (GI-711/5-1) within the priority program (SPP 1 736) Algorithms for Big Data.

Contents

Chapter 1

Introduction

1.1 Sparse and Low-Rank Modeling

William of Ockham (1287 - 1347) was an English Franciscan friar, theologian, and a main figure of *scholasticism*. Scholasticism was the predominant school of philosophy and teaching in medieval European universities between 1100 and 1700. It places a strong emphasis on the dialectical method. This method constitutes a disputation of a subject between two or more people who have different points of view, but have the common wish to establish the truth about the subject by means of reasonable argumentation. William of Ockham's most notable contribution to scholasticism was the idea that from multiple competing hypotheses with the same conclusion, one should select the one with the least assumptions. In other words, the simplest explanation should be preferred. This problem-solving principle has been named the *law of parsimony* or, in honor of the Franciscan friar, *Ockham's razor*. In his time, William of Ockham used the principle to advocate the existence of divine miracles. As a general and intuitive principle however, nowadays Ockham's razor is widely adopted across many domains and disciplines.

Machine learning is one such discipline. It is driven by technological progress and modern measurement methods, which have enabled the collection of vast amounts of data. However, usually insights are gained only after compression and reduction of the presumably complex data. Machine learning models that perform the tasks of compression and data reduction make use of Ockham's razor by assuming simple structures in the data. The two structures of interest in this thesis are *sparsity* and *low rank*. Here, sparse models have only a small fraction of parameters that are non-zero, and low-rank models exhibit a special type of sparsity that concerns the singular values of a matrix. A low-rank matrix has only few non-zero singular values. Since sparsity and low rank are central for this thesis, in the following we give a broad and conceptual overview on sparse and low-rank models. Afterwards, we outline the contributions of this thesis.

Sparse models. First, let us consider a typical example of sparse modeling: In a high-dimensional dataset, one may be interested in the selection of a small subset of features that are predictive for a target variable. For example, from among the numerous variables in a census dataset, one might identify age, education, and gender to be predictive for income. This reveals a central benefit of sparse models: Because of their small number of active features (non-zero parameters), they can be interpreted more easily by humans. In addition, learning sparse models often comes with the statistical benefit that fewer observations are required in order to learn the model parameters reliably. Besides, a reduced number of features also means less chance of overfitting.

The feature-selection problem above can be addressed by the *lasso*, which is a sparse regression model that was popularized by the highly influential work of Tibshirani [1996]. In its basic form, the lasso jointly minimizes a squared-error term and an ℓ_1-norm regularization term. Here, the squared-error term is the sum over the training samples of the squared differences between the actual and predicted (regressed) values of the target variable, respectively. Moreover, the ℓ_1-norm regularization term adds up the absolute values of the parameters, which induces sparsity on them. The lasso is theoretically well-founded and equipped with strong learning guarantees [Wainwright, 2009]. It poses a feature-selective alternative to ordinary least squares regression, which uses no regularization, and ridge regression [Tikhonov, 1943], which is based on ℓ_2-norm regularization, that is, Euclidean-norm regularization.

In regression models, only the interactions with one designated target variable are considered. More generally, one can be interested in the interactions between all the variables from a dataset. This yields *graphical* models that can serve more general queries. Graphical models are often represented using a (conditional) dependence graph. In this graph, the nodes represent variables, and the edges encode dependencies between these variables. Usually one assumes a sparse dependency graph, which results in sparse graphical models. They are often found in the natural sciences. For example, the *Ising* model of ferromagnetism in statistical mechanics [Ising, 1925] consists of binary variables that model the two states of magnetic dipole moments of atomic 'spins'. The spins are aligned in a grid structure. Hence, they interact only with a few neighbors, inducing an overall sparse graphical model structure.

Sparse graphical models can be estimated from data by the *graphical lasso* [Meinshausen and Bühlmann, 2006; Ravikumar et al., 2011; Jalali et al., 2011; Lee and Hastie, 2015]. The graphical lasso estimates a multivariate probability distribution that is parametrized by a symmetric matrix of pairwise interaction parameters. The objective function of the graphical lasso consists of two components: First, there is a likelihood term, which fits the distribution to the data. Second, similarly to the objective function of the lasso for sparse regression models, there is an ℓ_1-norm regularization term that induces sparsity on the matrix of pairwise interaction parameters. In this matrix, a zero entry indicates the absence of a dependence.

As a last model from the fast-growing sparse modeling literature we consider the *sparse coding* problem. In sparse coding, signals are represented as linear combinations of only a few elements (called atoms) from an over-complete basis/representation. For example, the atoms used for representing images can consist of small image patches with different edge patterns. Signal representations (sparse codes) for a given over-complete representation can be learned via *basis pursuit* [Chen et al., 2001]. The learnable parameters of basis pursuit are the coefficients of the atoms in the linear combination. Specifically, approximate basis pursuit minimizes a squared error, which makes sure that the signal is represented well by the linear combination of the atoms, and an additional ℓ_1-norm regularization term, which ensures that the coefficients are sparse. Interestingly, the objective function of basis pursuit formally coincides with the one of the lasso. Hence, basis pursuit and the lasso can be solved in the same way. This may be surprising since sparse coding and sparse regression are clearly not conceptually equivalent.

Low-rank models. Next, let us consider low-rank models. A typical task in low-rank modeling is matrix completion. A famous example is the Netflix challenge (Netflix prize), which took place between 2007 and 2009. The aim of the challenge was to predict user ratings for movies, where the only available data were previous ratings from the users. This data was given in form of an incomplete matrix that is indexed by the users and movies, respectively. Consequently, the task was to complete the matrix of user-movie ratings. Under the assumption that there are only a few prototypical user profiles from which individual profiles can be obtained as linear combinations, the matrix of user-movie ratings is low rank. In Fazel [2003], it was suggested to solve problems that involve low-rank matrices via convex optimization with nuclear-norm regularization. The nuclear norm can be understood as the ℓ_1-norm of the singular values. Consequently, it induces sparsity on the singular values, that is, it induces low rank on the matrix. For the matrix completion problem, a solution that involves convex optimization with nuclear-norm regularization was analyzed in [Candès and Recht, 2009].

Another important low-rank model that was introduced in [Pearson, 1901] is principal component analysis (PCA). It seeks to approximate a matrix by a low-rank matrix in the Frobenius-norm sense, that is, it minimizes the sum of the squared errors of the entries. The solution to PCA is the low-rank matrix that is constructed from the principal components whose corresponding singular values are of the largest magnitude. PCA is related to *factor analysis* [Spearman, 1904]. In contrast to PCA, factor models assume a small number of unobserved (latent) quantitative variables called *factors*. In factor models, interactions among the observed variables reflect indirect interactions due to the latent factors. As before, the interactions among the observed variables can be described by a matrix of pairwise interaction parameters. This matrix is low rank if there are only a few latent factors.

Estimating latent variables is especially important in the social sciences. This is because some quantities, such as, economic behavior or intelligence, do not allow for direct measurement. They represent latent constructs that require surrogate measurements. For example, personality traits are often measured by collecting answers of test takers to items of questionnaires. Here, the effect of the latent quantitative variables is modeled using *item response theory*, see [Embretson and Reise, 2013]. The items are commonly assumed to be conditionally independent given the typically small number of latent variables. Hence, the pairwise interaction parameter matrix of item response models is also low rank.

Sparse and low-rank models. Finally, we discuss models that have both sparse and low-rank components. The first model is motivated by the fact that principle component analysis is not robust with respect to gross data corruption. This brought up research on robust principle component analysis (RPCA). The first tractable definition of RPCA was independently introduced by Wright et al. [2009]; Candès et al. [2011], and Chandrasekaran et al. [2011]. They respectively decompose a data matrix into a low-rank and a sparse component, using convex optimization with the previously discussed nuclear-norm and ℓ_1-norm regularization techniques on the components. Here, as before, the low-rank component represents the principal components, whereas the sparse component accounts for the corrupted data. McCoy and Tropp [2011] and Xu et al. [2010] extended RPCA models to the setting where whole data points can be corrupted, that is, they go beyond the corruption of individual entries. In Chapter 2 of this thesis, we consider RPCA models with even more general data corruption mechanisms.

RPCA models with sparse and low-rank matrix decompositions inspired many subsequent works, where sparse and low-rank modeling coalesced [Chen et al., 2011; Sprechmann et al., 2015; Yu et al., 2017]. Of particular interest for this thesis are *fused latent and graphical models* that were first introduced by Chandrasekaran et al. [2012]. Fused latent and graphical models combine models with latent quantitative variables (factor and item response theory models) with graphical models. They can be useful because sparse graphical models may fail to account for spurious influences of non-observed quantities. Likewise, factor and item response theory models may fail to include direct dependencies between the observed variables. Fused latent and graphical models address these potential flaws by decomposing the pairwise interaction parameter matrix into direct and indirect interactions, where the indirect interactions are due to the latent variables. Typically, one assumes a small number of latent variables and that only a few of the observed variables interact directly. Hence, the parameter matrix is decomposed into a sparse and a low-rank component. In practice, such decompositions can be learned by solving a convex regularized likelihood optimization problem with the usual sparsity- and low-rank-inducing regularization terms. The analysis of this learning method for general fused latent and graphical models is the primary subject of Chapter 3 and of this thesis. In the following, we outline the contributions of this thesis.

1.2 Contributions

Robust principle component analysis. In many applications, data points consist of several measurements that can be naturally divided into groups. For example, the color of a pixel is described by the values from all color channels in multi-channel images. If a pixel is corrupted (or more generally the group of measurements), then likely all measurements from the group are corrupted. Hence, it makes sense to consider a data corruption mechanism that affects *groups* of measurements. This data corruption mechanism generalizes the previously considered mechanisms [Chandrasekaran et al., 2011; Candès et al., 2011; McCoy and Tropp, 2011; Xu et al., 2010].

In Chapter 2, which is based on the work [Nussbaum and Giesen, 2021], we show that the approach of decomposing a data matrix by means of convex optimization and regularization remains computationally tractable for the generalized data corruption mechanism. For that, the ℓ_1-norm is replaced by the $\ell_{1,2}$-norm, which is defined as the sum of the ℓ_2-norms of the groups that are prescribed by the assumed data corruption mechanism. Hence, the $\ell_{1,2}$-norm induces structured sparsity that is also called *group sparsity*. In summary, the generalized RPCA problem uses nuclear-norm and $\ell_{1,2}$-norm regularization to learn a decomposition of the corrupted data matrix into a low-rank and a group-sparse component.

We investigate when the generalized RPCA problem allows to exactly recover the components. Here, exact recovery can only be guaranteed if the low-rank and group-sparse components cannot be confused, that is, if the decomposition is *identifiable*. As Chandrasekaran et al. [2011] noticed, identifiability can be characterized by studying geometric objects, particularly *tangent spaces* to algebraic matrix varieties. We extend the characterization from sparse to group-sparse matrices. Afterwards, we provide deterministic and probabilistic conditions for exact recovery. We corroborate our theoretical findings with experiments on synthetic data. Moreover, experiments on several real-world datasets from different domains demonstrate the wide applicability of our generalized approach.

Fused latent and graphical models. The results in Chapter 3 are based on the seminal work by Chandrasekaran et al. [2012], who considered fused latent and graphical models with Gaussian observed variables. However, many real-world applications exhibit different types of variables. Particularly, discrete variables are important as they can represent arbitrary categories, states, or choices. This has laid the foundation to consider fused latent and graphical models for the more general class of pairwise *conditional Gaussian* distributions in this thesis. Conditional Gaussian distributions, see [Lauritzen, 1996], allow for observed quantitative *and* discrete variables. Their name is due to the fact that the quantitative variables always follow a Gaussian distribution when conditioned on a fixed set of values for the discrete variables.

Let us briefly discuss pairwise interactions when discrete variables are involved. Discrete variables can be equivalently represented by indicator variables for their outcomes, where an indicator variable of an outcome is set to one if the discrete variable takes on the value of the outcome, and is set to zero else. It is convenient to define pairwise interactions with discrete variables on their indicator variables. Therefore, the interaction of any variable with a discrete variable is described by several parameters that form a group. The complete absence of the interaction means that all interaction parameters from this group must be zero. Therefore, the sparse graphical modeling component of the interaction parameter matrix of fused latent and graphical models is group sparse in the presence of discrete variables.

The decomposition of the pairwise interaction parameter matrix of fused latent and graphical models can be estimated using a convex likelihood optimization problem with the usual group-sparsity and low-rank-inducing regularization. We show that under suitable identifiability conditions, this optimization problem allows to recover fused latent and graphical models *consistently* from data, that is, asymptotically and with high probability. We prove consistency in the *high-dimensional setting*, where the dependence on the number of observed variables, the number of latent variables, and the number of sample points is explicitly considered. The results in Chapter 3 are based on several published works that consider different types of observed variables, namely, models with observed binary variables in [Nussbaum and Giesen, 2019a], models with observed discrete variables in [Nussbaum and Giesen, 2020a], and models with observed binary and quantitative variables in [Nussbaum and Giesen, 2020b]. However, the consistency result for models with observed discrete and quantitative variables presented in this thesis is still more general than any of the published works so far.

Next, we support our theoretical findings with experiments on synthetic and real-world data from item response theory studies. These experiments are mostly borrowed from [Nussbaum and Giesen, 2020a]. Afterwards, we consider the problem of selecting suitable regularization parameters for the regularized likelihood optimization problem. Here, based on our work [Giesen et al., 2019b], we offer a principled alternative to *grid search* or *random search*. It builds upon *Benson's algorithm*, which approximates the set of all possible solutions that can be obtained by using different combinations of regularization parameters. Particularly, we introduce an adaptive variant of Benson's algorithm that is efficient and also works out of the box for a large class of optimization problems.

Altogether, many different optimization problems are considered in this thesis. To practically solve these problems, we have designed and implemented various solvers. These solvers are mostly based on the *alternating direction method of multipliers* (ADMM), see [Boyd et al., 2011]. The code for learning fused latent and graphical models has been made publicly available in form of an online Python code repository, see `https://github.com/franknu/cgmodsel`. This code repository also contains solvers for learning other conditional Gaussian distributions, for example, purely sparse graphical models. Hopefully, the available code and the overall work presented in this thesis will be found useful. Perhaps, they will even motivate further research at the intersection of sparse and low-rank modeling.

Chapter 2

Multi-View Robust Principal Component Analysis

2.1 Introduction

Principal component analysis (PCA) is a classical data dimension reduction technique based on the assumption that given high-dimensional data lies near some low-dimensional subspace, see [Pearson, 1901]. Formally, assume observed data points $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)} \in \mathbb{R}^m$ that are combined into a data matrix $\boldsymbol{X} \in \mathbb{R}^{m \times n}$. Approximating the data matrix $\boldsymbol{X}$ by a low-rank matrix $\boldsymbol{L}$ can be formulated as an optimization problem

$$\min_{\boldsymbol{L} \in \mathbb{R}^{m \times n}} \|\boldsymbol{X} - \boldsymbol{L}\| \quad \text{subject to} \quad \operatorname{rank}(\boldsymbol{L}) \leq k, \tag{2.1}$$

where $\|\cdot\|$ is some suitable norm. The classical and still popular choice, see [Hotelling, 1933; Eckart and Young, 1936], uses the Frobenius norm $\|\cdot\|_F$, which renders the optimization problem tractable. However, because of the squared penalty, the Frobenius norm does not perform well for *grossly corrupted data*. A single grossly corrupted entry in $\boldsymbol{X}$ can change the estimated low-rank matrix $\boldsymbol{L}$ significantly, that is, the Frobenius norm approach is not robust against data corruption. An obvious remedy is replacing the Frobenius norm by the ℓ_1-norm $\|\cdot\|_1$, but this renders the optimization problem intractable because of the non-convex rank constraint. An alternative way to achieve robustness is to explicitly model a component that captures data corruption. This leads to a decomposition of the data matrix

$$\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}$$

into a low-rank component $\boldsymbol{L}$ as before and a matrix $\boldsymbol{S}$ of outliers. The structure of the outlier matrix $\boldsymbol{S}$ depends on the data corruption mechanism and is commonly assumed to be sparse. In practice, low-rank + sparse decompositions can be computed efficiently through the convex optimization problem

$$\min_{\boldsymbol{L},\, \boldsymbol{S} \in \mathbb{R}^{m\times n}} \|\boldsymbol{L}\|_* + \gamma \|\boldsymbol{S}\|_{1,2} \quad \text{subject to} \quad \boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}, \tag{2.2}$$

where $\gamma > 0$ is a trade-off parameter. The nuclear norm $\|\cdot\|_*$ can be seen as a convex relaxation of a rank constraint and thus promotes low rank on $\boldsymbol{L}$, and the $\ell_{1,2}$-norm

$$\|\boldsymbol{S}\|_{1,2} = \sum_g \|\boldsymbol{s}_g\|_2$$

promotes *structured* sparsity on $\boldsymbol{S}$ given a partitioning of the entries into groups $\boldsymbol{s}_g$. The groups are determined by the assumed data corruption mechanism.

In the simplest data corruption mechanism, *individual* components of the data points can be corrupted. Under this model, the $\ell_{1,2}$-norm reduces to the ℓ_1-norm, that is, the groups $\boldsymbol{s}_g$ consist of single elements. Wright et al. [2009]; Candès et al. [2011], and Chandrasekaran et al. [2011] were first to investigate this model. They show that exact recovery using the specialized version of Problem (2.2) is possible, that is, in many cases the corruptions can be separated from the data.

An alternative corruption mechanism that was introduced independently by McCoy and Tropp [2011] and Xu et al. [2010] corrupts *whole* data points, which are referred to as outliers. Here, the groups $\boldsymbol{s}_g$ for the $\ell_{1,2}$-norm are the columns of the data matrix $\boldsymbol{X}$. Xu et al. [2010] showed that exact recovery is also possible in the column-sparse scenario, where only a few data points are outliers.

In this chapter, which is based on the work Nussbaum and Giesen [2021], we study a more general data corruption mechanism, where the data points are partitioned into d groups:

$$\boldsymbol{x}^{(i)} = (\boldsymbol{x}_1^{(i)}, \ldots, \boldsymbol{x}_d^{(i)}) \in \mathbb{R}^{m_1} \times \ldots \times \mathbb{R}^{m_d} = \mathbb{R}^m.$$

We assume that every group for each data point can be individually corrupted. Hence, the groups in Problem (2.2) are given by

$$\boldsymbol{S} = \left(\begin{array}{c|c|c|c} \boldsymbol{s}_{11} & \boldsymbol{s}_{12} & \cdots & \boldsymbol{s}_{1n} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \boldsymbol{s}_{d1} & \boldsymbol{s}_{d2} & \cdots & \boldsymbol{s}_{dn} \end{array}\right) \in \mathbb{R}^{m\times n} \quad \text{with} \quad \boldsymbol{s}_{ij} \in \mathbb{R}^{m_i}.$$

In Section 2.2, we show that exact recovery is still possible for our more general data corruption mechanism. This mechanism has a natural interpretation in terms of generalized *multi-view models* [Sun, 2013; Zhao et al., 2017; Zhang et al., 2019], where each data point is obtained by measurements from different sensors, and every sensor can measure several variables. Sensor failures in this model result in corrupted measurements for the group of variables measured by the failing sensor, but only for the data points that were measured while the sensor was not working correctly. Of course, data corruption and sensor failures are an abstraction for what can also be anomalies or outliers in applications, see Figure 2.1 for a real-world example.

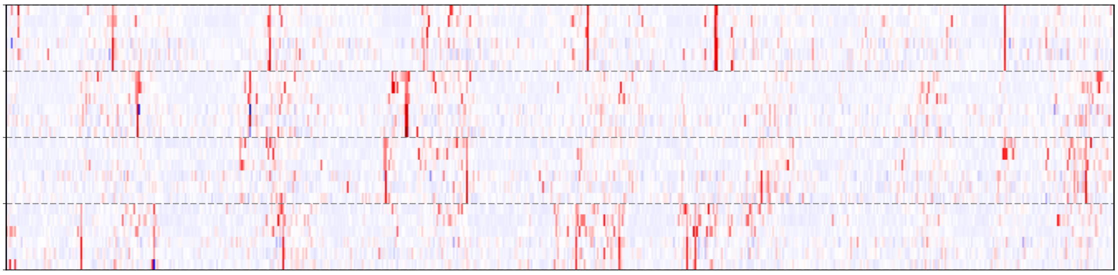

Figure 2.1: The electrical load profiles of four households from one week. Power consumption for each household is measured in terms of six quantities per time step (three phases of active and reactive power, respectively), which form groups of six elements. In the plot of the data, these groups respectively span six rows, whereas the columns represent time steps. Note that for this data, we expect data corruption (outliers) in the form of short-term usage of electrical devices. More details can be found in Section 2.4.

Note that the approach in [Candès et al., 2011; Chandrasekaran et al., 2011] corresponds to the special case where each sensor just measures a single variable, and the approach in [McCoy and Tropp, 2011; Xu et al., 2010] corresponds to the special case where a single sensor measures all variables. Of course, it is possible to think of data corruption mechanisms that correspond to even more general group structures, for example, rectangular groups formed by sub-matrices of the data matrix. The latter case does not pose any extra technical challenges. Hence, here we keep the exposition simple and stick with the generalized multi-view setting, calling the resulting models multi-view robust principle component analysis (MV-RPCA) models. These models are flexible and suitable for many real-world applications. We provide some examples in Section 2.4.

Some of the applications have data in the form of tensors. Therefore, we briefly discuss some additional important related work that concerns robust tensor principle component analysis (RTPCA). The most closely related works follow the convex optimization approach. Their main modeling effort lies in the generalization of low rank and the nuclear norm to tensors. For example, Huang et al. [2014] propose RTPCA using the sum of nuclear norms (SNN), which is based on Tucker rank. For 3D tensors, Zhang et al. [2014]; Lu et al. [2016, 2019]; Zhou and Feng [2017] use a nuclear norm based on t-SVD and tensor tubal rank. Most works, including [Huang et al., 2014; Lu et al., 2016, 2019], assume the simple data corruption mechanism that corrupts individual entries of the data tensor. [Zhang et al., 2014; Zhou and Feng, 2017] consider outliers distributed along slices of 3D tensors. The latter data corruption mechanism is a special case of our multi-view models when all groups have the same size (and the data matrix is viewed as a flattened version of a tensor). However, our general multi-view models allow for different group sizes, which gives them additional flexibility and distinguishes them from all existing RTPCA models.

2.2 Exact Recovery

In this section, we assume a data matrix $\boldsymbol{X}$ that has an underlying *true* decomposition $\boldsymbol{X} = \boldsymbol{L}^\star + \boldsymbol{S}^\star$ into a low-rank matrix $\boldsymbol{L}^\star$ and a group-sparse matrix $\boldsymbol{S}^\star$. We investigate under which conditions the pair $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ can be obtained as the guaranteed solution to Problem (2.2) with a suitably chosen regularization parameter γ. The outline of the analysis is as follows: In Section 2.2.1, before analyzing Problem (2.2) itself, we aim at answering the general question when low-rank and group-sparse matrix decompositions are *identifiable*. At the same time, we also contemplate Problem (2.2) on an intuitive level. We present the main result on exact recovery in Section 2.2.2, followed by some corollaries that concern the recovery of random decompositions in Section 2.2.3.

2.2.1 Identifiability and a non-convex problem version

In order to have a chance in separating the low-rank and group-sparse components in Problem (2.2), we must understand under which circumstances decompositions into low-rank and group-sparse matrices are *identifiable*. Here, identifiability of the decomposition means that it should not be possible to confuse the components. In this section, we formalize identifiability and derive conditions on the low-rank and group-sparse matrices that lead to identifiability.

For the study of identifiability, it turns out to be useful to consider Problem (2.2) as a convex relaxation of a problem that, instead of using nuclear-norm and $\ell_{1,2}$-norm regularization, constrains the component $\boldsymbol{L}$ to have a certain low rank and $\boldsymbol{S}$ to have a certain degree of group sparsity. These constraints can be expressed in terms of algebraic matrix varieties [Harris, 2013]. Here, we briefly introduce these varieties. First, the low-rank matrix variety of matrices with rank at most r is given by

$$\mathcal{L}(r) = \{\boldsymbol{L} \in \mathbb{R}^{m\times n} : \ \mathrm{rank}(\boldsymbol{L}) \le r\},$$

and second, the variety of group-structured matrices with at most s non-zero groups is given by

$$\mathcal{S}(s) = \{\boldsymbol{S} \in \mathbb{R}^{m\times n} : |\,\mathrm{gsupp}(\boldsymbol{S})| \le s\}.$$

Here,

$$\mathrm{gsupp}(\boldsymbol{S}) = \{(i,j) : 1 \le i \le d, 1 \le j \le n, \boldsymbol{s}_{ij} \not\equiv \boldsymbol{0}\}$$

is the *group support* of $\boldsymbol{S}$. Remember that $\boldsymbol{s}_{ij}$ is the sub-vector of the j-th column of $\boldsymbol{S}$ that corresponds to the i-th group of variables. Note that we assume the group structure to be fixed throughout, which is why the symbol $\mathcal{S}(s)$ of the group-sparse matrix variety does not include the dependency on the group structure.

In the sequel, we will discuss the identifiability of low-rank and group-sparse matrix decompositions based on the non-convex feasibility problem of finding $\boldsymbol{L}$ and $\boldsymbol{S}$ such that

$$\boldsymbol{L} \in \mathcal{L}(r), \;\; \boldsymbol{S} \in \mathcal{S}(s), \;\; \text{and} \;\; \boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S} \tag{2.3}$$

for given fixed r and s. The regularized Problem (2.2) can be seen as a convex relaxation of Problem (2.3). Note that Netrapalli et al. [2014] tried to directly solve a non-convex problem similar to Problem (2.3) for the non-group case, where individual entries can be corrupted. They assumed a priori estimates of rank and sparsity. Indeed, clearly the true decomposition $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ solves Problem (2.3) when the true varieties with $r = \operatorname{rank}(\boldsymbol{L}^\star)$ and $s = |\operatorname{gsupp}(\boldsymbol{S}^\star)|$ are used in Problem (2.3). However, these true varieties are unknown in practice. Nevertheless, the hypothetical Problem (2.3) provides valuable insights into the conditions that are necessary for successful recovery. This is because it allows us to define identifiability in the following sense: We call a decomposition $(\boldsymbol{L}, \boldsymbol{S}) \in \mathcal{L}(r) \times \mathcal{S}(s)$ identifiable in the product variety $\mathcal{L}(r) \times \mathcal{S}(s)$ if $(\boldsymbol{L}, \boldsymbol{S})$ uniquely solves Problem (2.3) with input $\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}$. Moreover, we call a pair $(\boldsymbol{L}, \boldsymbol{S}) \in \mathcal{L}(r) \times \mathcal{S}(s)$ *locally identifiable* if it is a locally unique solution to Problem (2.3), that is, if there exists some small ball such that it holds

$$(\boldsymbol{L} - \boldsymbol{\Delta}, \boldsymbol{S} + \boldsymbol{\Delta}) \notin \mathcal{L}(r) \times \mathcal{S}(s)$$

for all $\boldsymbol{\Delta} \neq \boldsymbol{0}$ from this small ball. We aim at a better understanding of local identifiability first. For that, observe that for determining local identifiability of a pair $(\boldsymbol{L}, \boldsymbol{S}) \in \mathcal{L}(r) \times \mathcal{S}(s)$, it is sufficient to consider points within the varieties that are close to $\boldsymbol{L}$ and $\boldsymbol{S}$, respectively. Hence, we need to characterize nearby points, which requires knowledge about the local geometry of the varieties. If $\boldsymbol{L} \in \mathcal{L}(r)$ has rank r and $\boldsymbol{S} \in \mathcal{S}(s)$ has s non-zero groups, then both are *smooth* points within their respective varieties. In this case, local geometry is determined by tangent spaces and local curvature. We discuss both for the respective varieties. First, the tangent space to the low-rank matrix variety $\mathcal{L}(r)$ at a rank-r matrix $\boldsymbol{L} \in \mathcal{L}(r)$ is given by

$$\mathcal{T}(\boldsymbol{L}) = \left\{\boldsymbol{U}\boldsymbol{X}^\mathsf{T} + \boldsymbol{Y}\boldsymbol{V}^\mathsf{T} : \; \boldsymbol{X} \in \mathbb{R}^{n\times r}, \boldsymbol{Y} \in \mathbb{R}^{m\times r}\right\},$$

where $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$ is the (restricted) singular value decomposition of $\boldsymbol{L}$ with $\boldsymbol{U} \in \mathbb{R}^{m\times r}$, $\boldsymbol{V} \in \mathbb{R}^{n\times r}$, and diagonal $\boldsymbol{D} \in \mathbb{R}^{r\times r}$. For a proof, see Appendix B.1. Second, the tangent space to the group-sparse matrix variety $\mathcal{S}(s)$ at $\boldsymbol{S}$ with $|\operatorname{gsupp}(\boldsymbol{S})| = s$ is given by

$$\mathcal{Q}(\boldsymbol{S}) = \{\boldsymbol{A} \in \mathbb{R}^{m\times n} : \; \operatorname{gsupp}(\boldsymbol{A}) \subseteq \operatorname{gsupp}(\boldsymbol{S})\}.$$

The group-sparse matrix variety has zero local curvature. Hence, if $\boldsymbol{S} + \boldsymbol{\Delta} \in \mathcal{S}(s)$ for small $\boldsymbol{\Delta}$, then it must hold $\boldsymbol{\Delta} \in \mathcal{Q}(\boldsymbol{S})$. However, in contrast to the group-sparse matrix variety, the low-rank matrix variety is locally curved. Because of the local curvature, if $\boldsymbol{L} - \boldsymbol{\Delta} \in \mathcal{L}(r)$ for small $\boldsymbol{\Delta}$, we can only conclude that $\boldsymbol{\Delta}$ must be a direction from some tangent space $\mathcal{T}(\boldsymbol{L}')$ to $\mathcal{L}(r)$ at a matrix $\boldsymbol{L}' \in \mathcal{L}(r)$ that is close to $\boldsymbol{L}$. Nevertheless, for local identifiability it suffices to only consider the tangent spaces $\mathcal{T}(\boldsymbol{L})$ and $\mathcal{Q}(\boldsymbol{S})$.

Lemma 2.1. *Let $\boldsymbol{L} \in \mathcal{L}(r)$ and $\boldsymbol{S} \in \mathcal{S}(s)$ be smooth points, that is, $\boldsymbol{L}$ has rank r and $\boldsymbol{S}$ has s non-zero groups. Assume that the tangent spaces $\mathcal{T}(\boldsymbol{L})$ and $\mathcal{Q}(\boldsymbol{S})$ are* transverse, *which means that*

$$\mathcal{T}(\boldsymbol{L}) \cap \mathcal{Q}(\boldsymbol{S}) = \{\boldsymbol{0}\}.$$

Then, the pair $(\boldsymbol{L}, \boldsymbol{S})$ is locally identifiable in $\mathcal{L}(r) \times \mathcal{S}(s)$.

To establish this result, one can prove that transversality of the tangent spaces extends to nearby tangent spaces, that is, one can show that the assumption $\mathcal{Q}(\boldsymbol{S}) \cap \mathcal{T}(\boldsymbol{L}) = \{\boldsymbol{0}\}$ also implies that $\mathcal{Q}(\boldsymbol{S}) \cap \mathcal{T}(\boldsymbol{L}') = \{\boldsymbol{0}\}$ as long as $\boldsymbol{L}'$ is sufficiently close to $\boldsymbol{L}$. This is precisely what the proof in Appendix B.2 does.

Now, considering only points $(\boldsymbol{L}, \boldsymbol{S}) \in \mathcal{L}(r) \times \mathcal{S}(s)$ with $\mathcal{T}(\boldsymbol{L}) \cap \mathcal{Q}(\boldsymbol{S}) = \{\boldsymbol{0}\}$ in our analysis leads to locally identifiable decompositions, though not globally identifiable. Let us consider a simple illustrative example for which we assume $n = d = 3$ and groups that consist of individual entries. For the example, assume that $\boldsymbol{L} = \boldsymbol{e}_3\boldsymbol{e}_3^\top \in \mathcal{L}(1)$ and let $\boldsymbol{S} = \boldsymbol{e}_1\boldsymbol{e}_1^\top \in \mathcal{S}(1)$, where $\boldsymbol{e}_i$ is the i-th standard basis vector. One can easily check that it holds $\mathcal{T}(\boldsymbol{L}) \cap \mathcal{Q}(\boldsymbol{S}) = \{\boldsymbol{0}\}$. Hence, the pair $(\boldsymbol{L}, \boldsymbol{S})$ is locally identifiable in $\mathcal{L}(1) \times \mathcal{S}(1)$, that is, the decomposition $\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}$ is locally unique around $(\boldsymbol{L}, \boldsymbol{S})$. However, exchanging the roles of $\boldsymbol{L}$ and $\boldsymbol{S}$ yields a different decomposition that is also in $\mathcal{L}(1) \times \mathcal{S}(1)$. Indeed, if the input matrix is $\boldsymbol{X} = \boldsymbol{e}_1\boldsymbol{e}_1^\top + \boldsymbol{e}_3\boldsymbol{e}_3^\top$, then Problem (2.3) and Problem (2.2) (with $\gamma = 1$) are both solved by $(\boldsymbol{e}_1\boldsymbol{e}_1^\top, \boldsymbol{e}_3\boldsymbol{e}_3^\top)$ and $(\boldsymbol{e}_3\boldsymbol{e}_3^\top, \boldsymbol{e}_1\boldsymbol{e}_1^\top)$, that is, their solutions are non-unique. This example shows that local identifiability is not yet enough to guarantee the unique recovery of the true components.

The problem in the example above is that *both* components are group sparse and low rank at the same time. Thus, they can easily be confused. It is much harder to find an alternative valid decomposition that still satisfies the group-sparsity and rank constraints in Problem (2.3) if the low-rank matrix is not (group) sparse since then taking away a sparse part likely increases rank. Similarly, if the group-sparse matrix is not low rank, then taking away a low-rank component is likely to increase the group support. In the following, we provide conditions to avoid that either component is simultaneously low rank and group sparse.

First, to ensure that $\boldsymbol{L}$ is not group sparse, we want its entries to be spread-out. This is the case if the row and column spaces of $\boldsymbol{L}$ are *incoherent*. Here, the incoherence of a subspace $V \subseteq \mathbb{R}^n$ is defined as $\operatorname{coh}(V) = \max_i \|P_V \boldsymbol{e}_i\|_2$, that is, the maximum length of the projection of a standard-basis vector $\boldsymbol{e}_i$ of $\mathbb{R}^n$ on the space V. Incoherence thus measures how well the subspace is aligned with the standard coordinate axes. A high value indicates that the subspace is well-aligned. For example, if V contains a standard basis vector, then it holds $\operatorname{coh}(V) = 1$. As noted by Chandrasekaran et al. [2011], in general it holds

$$\sqrt{\frac{k}{n}} \leq \operatorname{coh}(V) \leq 1$$

for a k-dimensional subspace. Now, we define the incoherence of $\boldsymbol{L}$ as

$$\operatorname{coh}(\boldsymbol{L}) = \max \left\{\operatorname{coh}(\operatorname{rowspace}(\boldsymbol{L})), \operatorname{coh}(\operatorname{colspace}(\boldsymbol{L}))\right\},$$

where $\operatorname{rowspace}(\boldsymbol{L})$ is the row space of $\boldsymbol{L}$ and $\operatorname{colspace}(\boldsymbol{L})$ is the column space of $\boldsymbol{L}$. We want $\operatorname{coh}(\boldsymbol{L})$ to be small since then the column respectively row vectors cannot be well-aligned with the respective standard basis vectors, which means that the entries must be spread-out. In this case, $\boldsymbol{L}$ is likely not group sparse.

Second, we define the maximum group degree $\operatorname{gdeg}_{\max}(\boldsymbol{S})$ as the maximum number of non-zero groups that appear in a row or column of $\boldsymbol{S}$. We want $\operatorname{gdeg}_{\max}(\boldsymbol{S})$ to be small since it implies that the non-zero groups are *not* concentrated in just a few rows and columns, which means that the matrix $\boldsymbol{S}$ is likely not low rank.

Having introduced these notions, the next lemma shows that bounding the product $\operatorname{coh}(\boldsymbol{L}) \operatorname{gdeg}_{\max}(\boldsymbol{S})$ implies transversality of the tangent spaces and thus local identifiability by Lemma 2.1.

Lemma 2.2. *Let $\boldsymbol{L} \in \mathcal{L}(r)$ and $\boldsymbol{S} \in \mathcal{S}(s)$ be smooth points as before. Let $\mathcal{T}(\boldsymbol{L})$ be the tangent space to the low-rank matrix variety at $\boldsymbol{L}$, and let $\mathcal{Q}(\boldsymbol{S})$ be the tangent space to the group-sparse matrix variety at $\boldsymbol{S}$. Define $\eta = \max_{i=1}^{d} m_i$ to be the maximum number of variables that a group spans. If*

$$\operatorname{coh}(\boldsymbol{L}) \operatorname{gdeg}_{\max}(\boldsymbol{S}) < \frac{1}{2\eta^{3/4}},$$

then the tangent spaces are transverse, that is, $\mathcal{T}(\boldsymbol{L}) \cap \mathcal{Q}(\boldsymbol{S}) = \{\boldsymbol{0}\}$.

Observe that for $(\boldsymbol{L}, \boldsymbol{S}) = (\boldsymbol{e}_3\boldsymbol{e}_3^\mathsf{T}, \boldsymbol{e}_1\boldsymbol{e}_1^\mathsf{T})$ from the previously discussed example it holds $\operatorname{coh}(\boldsymbol{L}) \operatorname{gdeg}_{\max}(\boldsymbol{S}) = 1$ such that the condition of Lemma 2.2 is not satisfied. In the next section, we will see that a slightly stronger upper bound on the product $\operatorname{coh}(\boldsymbol{L}^\star) \operatorname{gdeg}_{\max}(\boldsymbol{S}^\star)$ even allows the exact recovery of $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ by solving instances of Problem (2.2).

Below, we will use a stronger result to prove Lemma 2.2. For the stronger result we will formulate a weaker assumption that relies on more technical notions that measure how elements from different tangent spaces compare. Specifically, for smooth $\boldsymbol{L} \in \mathcal{L}(r)$ and $\boldsymbol{S} \in \mathcal{S}(s)$ with corresponding tangent spaces $\mathcal{T}(\boldsymbol{L})$ and $\mathcal{Q}(\boldsymbol{S})$, we define the following quantities:

$$\begin{aligned}\xi(\mathcal{T}(\boldsymbol{L})) &= \max_{\boldsymbol{M} \in \mathcal{T}(\boldsymbol{L}),\, \|\boldsymbol{M}\|=1} \|\boldsymbol{M}\|_{\infty,2} \quad \text{and} \\ \mu(\mathcal{Q}(\boldsymbol{S})) &= \max_{\boldsymbol{M} \in \mathcal{Q}(\boldsymbol{S}),\, \|\boldsymbol{M}\|_{\infty,2}=1} \|\boldsymbol{M}\|.\end{aligned}$$

These quantities are norm-compatibility constants of the $\ell_{\infty,2}$- and spectral norms when restricted to elements from different tangent spaces. Note that these norms are the dual norms of the regularizing norms that appear in Problem (2.2). Using these norms later facilitates the analysis of Problem (2.2). Also note that the definitions of the norm-compatibility constants generalize the ones in [Chandrasekaran et al., 2011] by replacing the non-group ℓ_∞-norm (maximum norm) by the group $\ell_{\infty,2}$-norm to account for the more general data corruption mechanism. The following lemma shows that incoherence and maximum group degree can be bounded from below in terms of the norm-compatibility constants, which will allow us to interpret these constants.

Lemma 2.3. *Let $\boldsymbol{L} \in \mathcal{L}(r)$ and $\boldsymbol{S} \in \mathcal{S}(s)$ be smooth points as before. Then, the following bounds hold:*

$$\operatorname{coh}(\boldsymbol{L}) \geq 1/2\eta^{-1/2}\,\xi(\mathcal{T}(\boldsymbol{L})) \quad \textit{and} \quad \operatorname{gdeg}_{\max}(\boldsymbol{S}) \geq \eta^{-1/4}\mu(\mathcal{Q}(\boldsymbol{S})).$$

Lemma 2.3, which is proven in Appendix B.3, helps to connect the intuition that we have about the incoherence of $\boldsymbol{L}$ and the maximum group degree of $\boldsymbol{S}$ to the technical norm-compatibility constants. Remember that we want $\operatorname{coh}(\boldsymbol{L})$ to be small to avoid confusion of $\boldsymbol{L}$ with a group-sparse matrix. Hence, $\xi(\mathcal{T}(\boldsymbol{L}))$ must be small since otherwise $\operatorname{coh}(\boldsymbol{L})$ is large because of the lower bound from Lemma 2.3. Similarly, we want $\operatorname{gdeg}_{\max}(\boldsymbol{S})$ to be small in order to ensure that $\boldsymbol{S}$ cannot be confused with a low-rank matrix. Therefore, $\mu(\mathcal{Q}(\boldsymbol{S}))$ must be small since otherwise $\operatorname{gdeg}_{\max}(\boldsymbol{S})$ is large because of the lower bound from Lemma 2.3. In summary, we want both $\xi(\mathcal{T}(\boldsymbol{L}))$ and $\mu(\mathcal{Q}(\boldsymbol{S}))$ to be small. This is reflected in the assumption of the following result that is stronger than Lemma 2.2.

Lemma 2.4. *Let $\boldsymbol{L} \in \mathcal{L}(r)$ and $\boldsymbol{S} \in \mathcal{S}(s)$ be smooth points as before. Suppose that it holds*

$$\xi(\mathcal{T}(\boldsymbol{L}))\mu(\mathcal{Q}(\boldsymbol{S})) < 1.$$

Then, the tangent spaces are transverse, that is, it holds $\mathcal{T}(\boldsymbol{L}) \cap \mathcal{Q}(\boldsymbol{S}) = \{\boldsymbol{0}\}$.

Proof. Let $\boldsymbol{0} \neq \boldsymbol{M} \in \mathcal{T}(\boldsymbol{L})$. We calculate

$$\begin{aligned}\|P_{\mathcal{Q}(\boldsymbol{S})}\boldsymbol{M}\| &\leq \mu(\mathcal{Q}(\boldsymbol{S}))\|P_{\mathcal{Q}(\boldsymbol{S})}\boldsymbol{M}\|_{\infty,2} \\ &\leq \mu(\mathcal{Q}(\boldsymbol{S}))\|\boldsymbol{M}\|_{\infty,2} \\ &\leq \mu(\mathcal{Q}(\boldsymbol{S}))\xi(\mathcal{T}(\boldsymbol{L}))\|\boldsymbol{M}\| < \|\boldsymbol{M}\|,\end{aligned}$$

where the first inequality uses the definition of $\mu(\mathcal{Q}(\boldsymbol{S}))$, the second inequality is easy (it also follows from the projection Lemma B.2 in Appendix B), the third inequality follows from $\boldsymbol{M} \in \mathcal{T}(\boldsymbol{L})$ and the definition of $\xi(\mathcal{T}(\boldsymbol{L}))$, and the last inequality follows from the assumption. It follows that $P_{\mathcal{Q}(\boldsymbol{S})}(\boldsymbol{M}) \neq \boldsymbol{M}$ such that $\boldsymbol{M}$ cannot be contained in $\mathcal{Q}(\boldsymbol{S})$. This implies transversality of the tangent spaces. ■

Lemma 2.2 can now be proven as a simple corollary of Lemma 2.4.

Proof of Lemma 2.2. It holds

$$\xi(\mathcal{T}(\boldsymbol{L}))\mu(\mathcal{Q}(\boldsymbol{S})) < 2\eta^{3/4}\operatorname{coh}(\boldsymbol{L})\operatorname{gdeg}_{\max}(\boldsymbol{S}) < 1,$$

where the first inequality follows from Lemma 2.3, and the second inequality from the assumption of Lemma 2.2. Hence, the claim follows from Lemma 2.4. ■

Let us now redirect our attention to the convex Problem (2.2), where we want to show that only slightly stronger assumptions than the ones from Lemma 2.2 and Lemma 2.4 for the true decomposition $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ allow guaranteed exact recovery of $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ by Problem (2.2) with input $\boldsymbol{X} = \boldsymbol{L}^\star + \boldsymbol{S}^\star$.

2.2.2 Main results on unique and exact recovery

Before we present the main results, we state a sufficient condition that warants that $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ is the unique solution to Problem (2.2) (with suitably chosen γ). This condition is based on the first-order optimality conditions of Problem (2.2), which any solution to Problem (2.2) must satisfy. We derive it from the Lagrangian

$$\mathcal{L}(\boldsymbol{L}, \boldsymbol{S}, \boldsymbol{Z}) = \|\boldsymbol{L}\|_* + \gamma\|\boldsymbol{S}\|_{1,2} + \langle \boldsymbol{Z}, \boldsymbol{X} - \boldsymbol{L} - \boldsymbol{S}\rangle,$$

where $\boldsymbol{Z}$ are the dual variables for the constraint $\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}$, and $\langle\cdot,\cdot\rangle$ denotes the standard scalar product on matrices. The first-order optimality conditions with respect to $\boldsymbol{L}$ and $\boldsymbol{S}$ require that $\boldsymbol{Z}$ is a subgradient from the $\ell_{1,2}$-norm and the nuclear norm *subdifferentials*, that is, it must hold $\boldsymbol{Z} \in \partial\|\boldsymbol{L}\|_*$ and $\boldsymbol{Z} \in \gamma\partial\|\boldsymbol{S}\|_{1,2}$. The norm subdifferentials can be characterized using dual norms, see [Watson, 1992] and Lemma C.13 in Appendix C.3.5. First, it holds $\boldsymbol{Z} \in \partial\|\boldsymbol{L}\|_*$ if and only if

$$P_{\mathcal{T}(\boldsymbol{L})}(\boldsymbol{Z}) = \boldsymbol{U}\boldsymbol{V}^\mathsf{T} \quad \text{and} \quad \|P_{\mathcal{T}(\boldsymbol{L})^\perp}(\boldsymbol{Z})\| \le 1,$$

where $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$ is a singular value decomposition of $\boldsymbol{L}$, and $\|\cdot\|$ denotes the spectral norm, which is dual to the nuclear norm. Next, it holds $\boldsymbol{Z} \in \gamma\partial\|\boldsymbol{S}\|_{1,2}$ if and only if

$$P_{\mathcal{Q}(\boldsymbol{S})}(\boldsymbol{Z}) = \gamma\operatorname{gsign}(\boldsymbol{S}) \quad \text{and} \quad \|P_{\mathcal{Q}^\perp(\boldsymbol{S})}(\boldsymbol{Z})\|_{\infty,2} \le \gamma,$$

where the *group-sign* function maps a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ onto the matrix $\operatorname{gsign}(\boldsymbol{A}) \in \mathbb{R}^{m\times n}$ with

$$\operatorname{gsign}(\boldsymbol{A})_{ij} = \begin{cases} \boldsymbol{a}_{ij}/\|\boldsymbol{a}_{ij}\|_2, & \boldsymbol{a}_{ij} \neq \boldsymbol{0} \\ \boldsymbol{0}, & \text{else} \end{cases}, \qquad i \in [d], j \in [n].$$

Note that $\mathcal{Q}^\perp(\boldsymbol{S})$ are the matrices with complementary group support.

Based on the first-order optimality conditions, the following result states that $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ uniquely solves Problem (2.2) provided that the tangent spaces $\mathcal{T}(\boldsymbol{L}^\star)$ and $\mathcal{Q}(\boldsymbol{S}^\star)$ are transverse and given a dual $\boldsymbol{Z}$ that *strictly* satisfies the subgradient conditions above.

Proposition 2.5. *Suppose that $\boldsymbol{X} = \boldsymbol{L}^\star + \boldsymbol{S}^\star$. Then, $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ is the unique minimizer of Problem (2.2) if the following conditions are satisfied:*

1. *The tangent spaces are transverse, that is, it holds $\mathcal{T}(\boldsymbol{L}^\star) \cap \mathcal{Q}(\boldsymbol{S}^\star) = \{\boldsymbol{0}\}$.*
2. *There exists a subgradient $\boldsymbol{Z} \in \partial\|\boldsymbol{L}^\star\|_* \cap \gamma\partial\|\boldsymbol{S}^\star\|_{1,2}$ that satisfies the strict dual-feasible conditions*

$$\|P_{\mathcal{T}^\perp(\boldsymbol{L}^\star)}(\boldsymbol{Z})\| < 1 \quad \textit{and} \quad \|P_{\mathcal{Q}^\perp(\boldsymbol{S}^\star)}(\boldsymbol{Z})\|_{\infty,2} < \gamma.$$

The first condition ensures uniqueness in tangential directions. The second condition intuitively ensures uniqueness in normal directions: When one component is perturbed in a normal direction, then the corresponding subgradient changes in a non-continuous way because of the strict inequalities.

The idea for the proof of uniqueness, which can be found in Appendix B.4.1, is to assume the existence of another minimizer $(\boldsymbol{L}^\star - \boldsymbol{\Delta}, \boldsymbol{S}^\star + \boldsymbol{\Delta})$ with the goal of showing that $\boldsymbol{\Delta} = \boldsymbol{0}$. For that, the proof uses the subgradient property and the subgradient characterizations to show that $\boldsymbol{\Delta} \in \mathcal{T}(\boldsymbol{L}^\star) \cap \mathcal{Q}(\boldsymbol{S}^\star)$. This implies that $\boldsymbol{L}^\star - \boldsymbol{\Delta} \in \mathcal{T}(\boldsymbol{L}^\star)$ and $\boldsymbol{S}^\star + \boldsymbol{\Delta} \in \mathcal{Q}(\boldsymbol{S}^\star)$. However, from the transversality $\mathcal{T}(\boldsymbol{L}^\star) \cap \mathcal{Q}(\boldsymbol{S}^\star) = \{\boldsymbol{0}\}$ it follows that $\boldsymbol{\Delta}$ must be zero.

Proposition 2.5 requires the existence of a strictly dual feasible $\boldsymbol{Z}$, where the strict dual feasibility conditions depend on the regularization parameter γ. We now present the main result, which determines a range of values for γ for which a strictly dual feasible $\boldsymbol{Z}$ exists. The main result makes an assumption that is only slightly stronger than the one of Lemma 2.4.

Theorem 2.6. *Suppose that $\boldsymbol{X} = \boldsymbol{L}^\star + \boldsymbol{S}^\star$. If*

$$\xi(\mathcal{T}(\boldsymbol{L}^\star))\mu(\mathcal{Q}(\boldsymbol{S}^\star)) < 1/6,$$

then the range

$$(\gamma_{\min}, \gamma_{\max}) = \left(\frac{\xi(\mathcal{T}(\boldsymbol{L}^\star))}{1 - 4\xi(\mathcal{T}(\boldsymbol{L}^\star))\mu(\mathcal{Q}(\boldsymbol{S}^\star))}, \frac{1 - 3\xi(\mathcal{T}(\boldsymbol{L}^\star))\mu(\mathcal{Q}(\boldsymbol{S}^\star))}{\mu(\mathcal{Q}(\boldsymbol{S}^\star))}\right)$$

is non-empty. Moreover, for any γ in that range, Problem (2.2) with regularization parameter γ is uniquely solved by $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$.

We prove this theorem in Appendix B.4.2, where the goal is to show the existence of a dual variable $\boldsymbol{Z}$ as required by Proposition 2.5. In conjunction with Lemma 2.3, Theorem 2.6 yields an immediate corollary.

Corollary 2.7. *Suppose that $\boldsymbol{X} = \boldsymbol{L}^\star + \boldsymbol{S}^\star$. Let*

$$\operatorname{coh}(\boldsymbol{L}^\star)\operatorname{gdeg}_{\max}(\boldsymbol{S}^\star) < 1/12\,\eta^{-3/4}.$$

Then, the interval

$$(\gamma^\circ_{\min}, \gamma^\circ_{\max}) = \left(\frac{2\eta^{1/2}\ \operatorname{coh}(\boldsymbol{L}^\star)}{1-8\eta^{3/4}\operatorname{coh}(\boldsymbol{L}^\star)\operatorname{gdeg}_{\max}(\boldsymbol{S}^\star)}, \frac{1-6\eta^{3/4}\operatorname{coh}(\boldsymbol{L}^\star)\operatorname{gdeg}_{\max}(\boldsymbol{S}^\star)}{\eta^{1/4}\operatorname{gdeg}_{\max}(\boldsymbol{S}^\star)}\right)$$

is non-empty. Moreover, for any γ in that range, Problem (2.2) *with regularization parameter γ has the* unique *solution $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$.*

Proof. This is straightforward using the lower bounds on $\operatorname{gdeg}_{\max}(\boldsymbol{S}^\star)$ and $\operatorname{coh}(\boldsymbol{L}^\star)$ from Lemma 2.3, which imply that

$$\xi(\mathcal{T}(\boldsymbol{L}^\star))\mu(\mathcal{Q}(\boldsymbol{S}^\star)) \le \eta^{\frac{1}{4}}\operatorname{coh}(\boldsymbol{L}^\star)\operatorname{gdeg}_{\max}(\boldsymbol{S}^\star)2\eta^{\frac{1}{2}} < \frac{1}{6},$$

where the last inequality follows from the assumption. Hence, we can apply Theorem 2.6. One can check by plugging in the lower bounds from Lemma 2.3 that the range $(\gamma^\circ_{\min}, \gamma^\circ_{\max})$ of values for γ is a non-empty sub-range of the range $(\gamma_{\min}, \gamma_{\max})$ given in Theorem 2.6. ■

It should be noted that in real-world situations the true maximum group degree and incoherence are unknown. This leaves the choice of γ up to the user. In the experimental Section 2.4.1, we investigate two heuristics for selecting the regularization parameter γ. In preparation for the experiments, we consider random low-rank and group-sparse decompositions in the next section.

2.2.3 Random decompositions

Since we intend to experiment also with synthetic data, we need to generate *random* low-rank + group-sparse decompositions. Therefore, we introduce a random decomposition model next, and we provide a theoretical result that concerns the recovery of random decompositions drawn from this model.

As in [Candès and Recht, 2009], we assume that a rank-r-matrix $\boldsymbol{L}^\star$ is drawn from the *random orthogonal model*, that is, by setting $\boldsymbol{L}^\star = \boldsymbol{U}\boldsymbol{V}^\top/\sqrt{mn}$, where $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times r}$ are drawn at random with independent standard Gaussian entries. The column spaces of $\boldsymbol{U}$ and $\boldsymbol{V}$ are incoherent with high probability. Indeed, by [Candès and Recht, 2009, Lemma 2.2], there exists a constant c such that it holds

$$\begin{aligned}\operatorname{coh}(\boldsymbol{L}^\star) &= \max\{\operatorname{coh}(\operatorname{colspace}(\boldsymbol{U})), \operatorname{coh}(\operatorname{colspace}(\boldsymbol{V}))\} \\ &\le c\max\left\{\frac{\max(r, \log m)}{m}, \frac{\max(r,\log n)}{n}\right\}\end{aligned} \tag{2.4}$$

with high probability, that is, with a probability that converges to one as m and n grow to infinity.

Next, we sample $\boldsymbol{S}^\star$ as follows: First, the group support $\operatorname{gsupp}(\boldsymbol{S}^\star)$ is sampled at random using independent Bernoulli variables, where each group is non-zero with probability p. Note that this type of sampling is characteristic for $G(n,p)$ random graph models, see for example [Bollobás, 2001]. As in [Candès et al., 2011], we sample the entries of the groups that belong to the support uniformly at random from $\{-1, 1\}$. Under this random group sparsity model, the maximum group degree is independent from the precise values of the non-zero entries. Specifically, the following holds:

Lemma 2.8. *Let $a = \max\{n, d\}$. If $\boldsymbol{S}^\star$ is drawn from the random group sparsity model, then the maximum group degree satisfies with high probability (that converges to one as n and d grow to infinity) that*

$$\operatorname{gdeg}_{\max}(\boldsymbol{S}^\star) \leq 2ap + 3\sqrt{ap}.$$

Proof. To bound the maximum group degree, we must bound the number of non-zero groups in each row and column. We bound the number of non-zero groups for a single row first. The result then follows from applying a union bound.

The number of non-zero groups in a fixed row is a binomially-distributed random variable $Z \sim \operatorname{Bin}(n,p)$. A consequence of Talagrand's inequality is that for $0 \leq t \leq np = \mathbb{E}Z$ it holds

$$\mathbb{P}(Z \geq np + t + 3\sqrt{np}) \leq \exp\left(-t^2/(16np)\right),$$

see [Habib et al., 2013]. We set $t = np$ and obtain

$$\mathbb{P}(Z \geq 2np + 3\sqrt{np}) \leq \exp\left(-np/16\right).$$

Similarly, we have for the columns that

$$\mathbb{P}(Z \geq 2dp + 3\sqrt{dp}) \leq \exp\left(-dp/16\right).$$

Hence, with $a = \max(n, d)$ and by the union bound, the probability that any row or column has more than $2ap + 3\sqrt{ap}$ non-zero groups is at most $m\exp\left(-np/16\right) + n\exp\left(-dp/16\right)$, which is small for (comparably) large n and d. ■

The following corollary shows that if the group-selection probability p is not too high, then random decompositions can be recovered exactly with high probability.

Corollary 2.9. *Let $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ be sampled from the random decomposition model (with sufficiently large n and d). Let*

$$p < \frac{\left(\sqrt{9 + 2/3\eta^{-3/4}/\kappa} - 3\right)^2}{16a},$$

where

$$\kappa = c \max \left\{ \frac{\max(r, \log m)}{m}, \frac{\max(r, \log n)}{n} \right\}$$

is as in Inequality (2.4). *Then, the assumption of Corollary 2.7 holds with high probability. Hence, with high probability, the components* $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ *are the guaranteed solution to Problem* (2.2) *with input* $\boldsymbol{X} = \boldsymbol{L}^\star + \boldsymbol{S}^\star$ *and* $\gamma \in (\gamma^\circ_{\min}, \gamma^\circ_{\max})$.

Proof. We show that the assumption

$$\operatorname{coh}(\boldsymbol{L}^\star) \operatorname{gdeg}_{\max}(\boldsymbol{S}^\star) < 1/12\, \eta^{-3/4}$$

of Corollary 2.7 holds with high probability. Using the upper bound on the maximum group degree from Lemma 2.8 and that by Inequality (2.4) the incoherence satisfies $\operatorname{coh}(\boldsymbol{L}^\star) \leq \kappa$ with high probability, it suffices to show that

$$(2ap + 3\sqrt{ap})\, \kappa < 1/12\, \eta^{-3/4}.$$

This is equivalent to

$$p + \frac{3}{2\sqrt{a}} \sqrt{p} - \frac{1}{24a}\, \eta^{-3/4} \kappa^{-1} < 0,$$

which is a quadratic inequality in $\sqrt{p}$. Solving it yields

$$\sqrt{p} < \frac{-3}{4\sqrt{a}} + \sqrt{\frac{9}{16a} + \frac{1}{24a} \eta^{-3/4} \kappa^{-1}}.$$

Taking squares, it follows that

$$p < \frac{\left(\sqrt{9 + 2/3 \eta^{-3/4} \kappa^{-1}} - 3 \right)^2}{16a}.$$

This finishes the proof after applying Corollary 2.7. ■

Note that the right-hand side of the inequality in Corollary 2.9 becomes small if the rank r is large. Hence, for large rank r, the group-selection probability p is required to be small in order to guarantee exact recovery with high probability. Moreover, if r and p are both small, then exact recovery should be easy.

2.3 ADMM Algorithm

Similar to [Candès et al., 2011], we derive an *alternating direction method of multipliers* (ADMM) algorithm for Problem (2.2). This problem has augmented Lagrangian

$$\begin{aligned}\mathcal{L}(\boldsymbol{S},\boldsymbol{L},\boldsymbol{Z}) &= \gamma\|\boldsymbol{S}\|_{1,2} + \|\boldsymbol{L}\|_* + \langle \boldsymbol{Z}, \boldsymbol{X}-\boldsymbol{S}-\boldsymbol{L}\rangle + \frac{1}{2\kappa}\|\boldsymbol{X}-\boldsymbol{S}-\boldsymbol{L}\|_F^2 \\ &= \gamma\|\boldsymbol{S}\|_{1,2} + \|\boldsymbol{L}\|_* + \frac{1}{2\kappa}\|\boldsymbol{X}-\boldsymbol{S}-\boldsymbol{L}+\kappa\boldsymbol{Z}\|_F^2 - \frac{\kappa}{2}\|\boldsymbol{Z}\|_F^2,\end{aligned}$$

where $\boldsymbol{Z}$ are the dual variables for the constraint $\boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}$ and $\kappa > 0$. Minimization of the augmented Lagrangian w.r.t. $\boldsymbol{S}$ and $\boldsymbol{L}$ is equivalent to solving proximal operators with known solutions. Consequently, after initialization, ADMM performs the following updates:

$$\begin{cases}\boldsymbol{S}^{k+1} & = \arg\min_{\boldsymbol{S}} \mathcal{L}(\boldsymbol{S},\boldsymbol{L}^k,\boldsymbol{Z}^k) \\ & = \mathrm{gShrink}(\boldsymbol{X}-\boldsymbol{L}^k+\kappa\boldsymbol{Z}^k,\gamma\kappa), \\ \boldsymbol{L}^{k+1} & = \arg\min_{\boldsymbol{L}} \mathcal{L}(\boldsymbol{S}^{k+1},\boldsymbol{L},\boldsymbol{Z}^k) \\ & = \mathrm{sShrink}(\boldsymbol{X}-\boldsymbol{S}^{k+1}+\kappa\boldsymbol{Z}^k,\kappa), \\ \boldsymbol{Z}^{k+1} & = \boldsymbol{Z}^k + \kappa^{-1}(\boldsymbol{X}-\boldsymbol{S}^{k+1}-\boldsymbol{L}^{k+1}).\end{cases}$$

Here, the group soft-shrinkage operation acts on the (i,j)-th group as

$$[\mathrm{gShrink}(\boldsymbol{Z},\kappa)]_{ij} = \boldsymbol{z}_{ij}\cdot\max\left\{1-\frac{\kappa}{\|\boldsymbol{z}_{ij}\|_2},0\right\}. \tag{2.5}$$

Remember that $\boldsymbol{z}_{ij}$ is the sub-vector that corresponds to the i-th group of variables in the j-th column of $\boldsymbol{Z}$. Moreover, the spectral shrinkage operator is given by

$$\mathrm{sShrink}(\boldsymbol{Z},\kappa) = \boldsymbol{U}\,\mathrm{Shrink}(\boldsymbol{E},\kappa)\boldsymbol{V}^\mathsf{T},$$

where $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{E}\boldsymbol{V}^\mathsf{T}$ is the singular value decomposition of $\boldsymbol{Z}$ and

$$\mathrm{Shrink}(\boldsymbol{z},\kappa) = \mathrm{sign}(\boldsymbol{z})\ \max\{\boldsymbol{z}-\kappa,\}.$$

ADMM is known to converge under mild assumptions and we use the termination criteria from [Boyd et al., 2011]. The main computational burden of the presented algorithm lies in the computation of all singular values that are greater than the threshold κ, along with their left and right singular vectors. To accelerate our solver, we performed this task using fast randomized singular value thresholding based on [Halko et al., 2011]. While this can be quite efficient, it only provides an approximate solution. This means that it is not clear whether ADMM still converges. As long as the number of singular values that must be computed is not too high, our experiments indicate that convergence is not harmed. However, for a generic solver that can deal with large scale problems it would be desirable to guarantee its robustness for all kinds of input data.

2.4 Experiments

In this section, we perform experiments with synthetic and real-world data for robust principle component analysis. We use the ADMM algorithm from Section 2.3 for solving Problem (2.2).

2.4.1 Synthetic data

In our first experiment, we intent to experimentally verify the theory from Section 2.2. For that, we generate synthetic data in the form of random pairs $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ that we sample according to the random decomposition model that we introduced in Section 2.2.3. For each random decomposition $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$, we check if Problem (2.2) can be used to exactly recover $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$, using only the compound matrix $\boldsymbol{L}^\star + \boldsymbol{S}^\star$ as input. Here, our main goal is to vary the rank r of $\boldsymbol{L}^\star$ and the group-selection probability p for $\boldsymbol{S}^\star$, where as a consequence of Corollary 2.9, we expect that successful recovery is more likely possible if $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ is sampled with not too large r and p.

More specifically for this experiment, we fix $n = 500$ variables and the group structure $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{100}) \in \mathbb{R}^{500}$, where each group $\boldsymbol{x}_i \in \mathbb{R}^5$ consists of five features. Hence, $\boldsymbol{X} = \boldsymbol{L}^\star + \boldsymbol{S}^\star \in \mathbb{R}^{500\times 500}$. Then, for selected pairs (r, p), we respectively create 10 different random decompositions to average out sampling effects. We try to recover these decompositions by solving instances of Problem (2.2). However, we still need to choose a suitable regularization parameter γ for each problem. In the following, we compare the rates of successful recovery of two different heuristics for choosing γ.

For the first heuristic, observe that according to Theorem 2.7 exact recovery is possible for a *range* of values for γ. Hence, if successful recovery is possible for a problem, then we expect that there exists an interval of regularization parameters that respectively yield the correct solution. In particular, the solution is the same for all γ from this interval: We say that the solution is *stable* in this interval. As in [Chandrasekaran et al., 2011], we use this fact to search for an interval of values for the regularization parameter γ, where the solution to Problem (2.2) is stable (and both components are non-zero). If the search for such an interval is successful, then we check if the solution, which is the same for all γ from the interval, has the correct algebraic properties. If this is the case, we consider the recovery for the given problem as successful. Otherwise, we declare failure.

For convenience, we rewrite the objective of Problem (2.2) as $(1-\alpha)\|\boldsymbol{L}\|_* + \alpha\|\boldsymbol{S}\|_{1,2}$ and denote its solution by $(\boldsymbol{L}_\alpha, \boldsymbol{S}_\alpha)$, where α is in the *compact* interval $[0, 1]$. Then, we equivalently search for an interval of values for α, where the solution does not change. For that, we track how the solution changes by calculating the differences

$$\mathrm{diff}_\alpha = \|\boldsymbol{L}_{\alpha-\delta} - \boldsymbol{L}_\alpha\|_F + \|\boldsymbol{S}_{\alpha-\delta} - \boldsymbol{S}_\alpha\|_F$$

along the solution path obtained from a grid search with step size $\delta = 10^{-2}$, see Figure 2.2. The change of the solution follows a typical pattern, which can also be seen in Figure 2.2. There are three intervals, where the solution is stable: First, for very small values of α, there is little group-sparse regularization, hence the solution has a zero low-rank component. Likewise, for too large values of α, the solution always has a zero group-sparse component. The third interval with a stable solution in the middle is the one that we are looking for. In the example shown in Figure 2.2, the recovery error

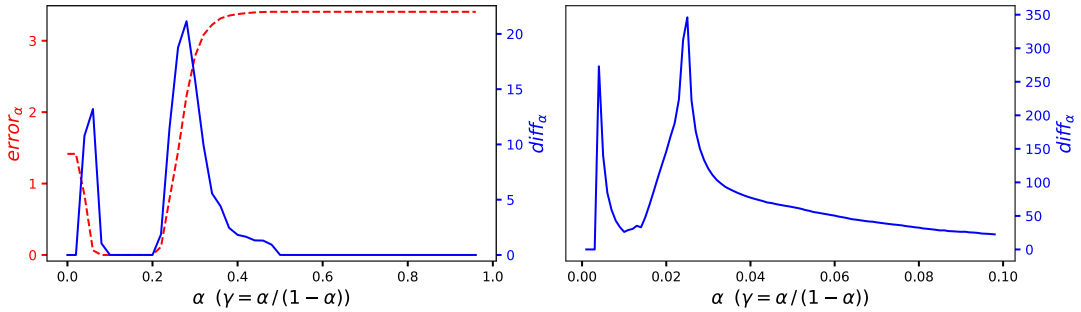

Figure 2.2: Search for a stable solution to Problem (2.2). The blue line respectively shows the change of the solution at each step of the grid search. *Left (synthetic data, step size* $\delta = 10^{-2}$*)*: The additional red line shows the recovery error, which is unknown in practice. For roughly $\alpha \in [0.1, 0.2]$, the solution is stable with almost zero recovery error. Hence, for this range, the solution is correct. *Right (loadprofiles real-world data, step size* $\delta = 10^{-3}$*)*: For roughly $\alpha \in [0.01, 0.015]$ the solution does not change much. The change is not completely zero. Still, the solution is relatively stable, particularly, the same structural discoveries (a low-rank day-and-night pattern) can be made for all values of α in the stable range. Note that the change *diff*$_\alpha$ is larger for the loadprofiles dataset because its dimensions are also larger.

$$\mathrm{error}_\alpha = \|\boldsymbol{L}_\alpha - \boldsymbol{L}^\star\|_F + \|\boldsymbol{S}_\alpha - \boldsymbol{S}^\star\|_F$$

is close to zero for all values in this interval. Note that the recovery error is unknown in practice.

The search for γ (or equivalently α) as outlined above requires solving several instances of Problem (2.2). Therefore, as a second heuristic, we also compare the rate of successful recovery to the rate when the ad-hoc choice $\gamma = 1/\sqrt{\max(m,n)}$ is used instead of searching. This value was suggested for learning RPCA decompositions under entry-wise data corruption, see [Candès et al., 2011].

The results of the experiment, see Figure 2.3, support the theory and effectively demonstrate that exact recovery is possible. Moreover, they confirm that for smaller r and p, that is, for smaller ranks and group-selection probabilities, successful recovery becomes easier.

Comparison of the first two plots in Figure 2.3 shows that exploiting prior knowledge about corruptions leads to improved results: The area for successful recovery

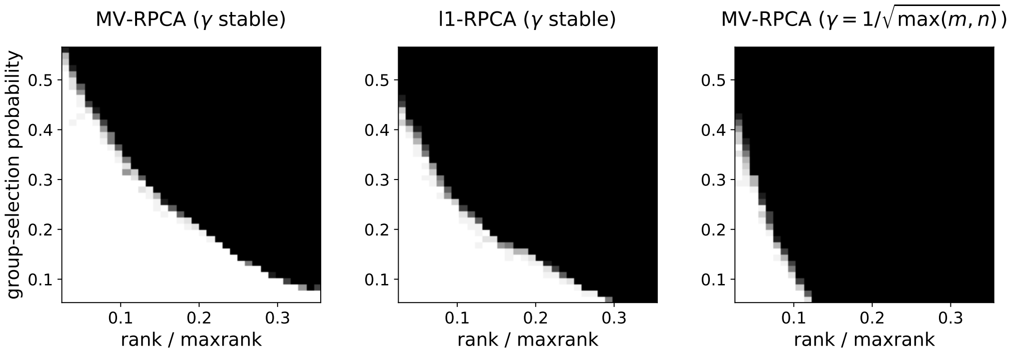

Figure 2.3: Recovery results for varying rank r (displayed as fractions $r/\max(m,n)$ of the maximum possible rank) and varying group-selection probability p. Trials were repeated 10 times for selected pairs (r,p). Empirical success probabilities are encoded as grey values, where white indicates a probability of 1 and black a probability of 0. The plots show from left to right: (a) the results when γ is selected based on a search for a stable solution, (b) the results when the ℓ_1-norm is used instead of the $\ell_{1,2}$ group-norm for regularization, and (c) the results that correspond to the ad-hoc choice $\gamma = 1/\sqrt{\max(m,n)}$.

is larger for multi-view robust principle component analysis (MV-RPCA) compared standard robust PCA with ℓ_1-norm regularization (l1-RPCA). This is because intuitively it easier to find a group of corrupted entries than to find each entry of the corrupted group individually.

Finally, comparing the first and last plot in Figure 2.3 it turns out that it may also pay off to perform the search for an interval, where the solution is stable. This is because decompositions with much greater ranks and group-selection probabilities can still be recovered successfully, though using the ad-hoc choice fails. On the other hand, the ad-choice can be tuned by hand if there is a priori knowledge about the solution. For example, if the outlier matrix is very sparse, then a larger value of γ can be used.

2.4.2 Real-world data

In the following, our goal is to demonstrate the wide applicability of robust principal component analysis for generalized multi-view models that have group-structured observations. For that, we briefly discuss four real-world applications.

Identification of periods of power consumption. In households, aside from the base load, power consumption usually takes place infrequently and during a limited period of time when electrical devices are turned on. Thus, momentary power consumption in households has characteristic features of outliers. Hence, we aim at showing that our model can be used to identify periods of large power consumption from electrical grid data. Specifically, we use a dataset that contains the electrical load profiles of 74 representative German residential buildings from the year 2010. The dataset, which was obtained from [Tjaden et al., 2015], constitutes a time series with a temporal resolution of one second. For illustrative purposes, we restrict the dataset to the first week. For each residential building, the electrical load profile consists of six quantities that correspond to three phases, respectively, of active and reactive power. Hence, each of the 74 residential buildings entails a group of six elements. Thus, at each time step a 444-dimensional vector is observed. In total, the data matrix $\boldsymbol{X}$ is of size $444 \times 10\,080$, including one observation for each second of the week. Note that sample data from the first four households is shown in Figure 2.1.

The solution to Problem (2.2) is stable around $\gamma = 10^{-2}$. Figure 2.4 shows the corresponding decomposition. There is a noticeable general pattern of electrical load profiles that is explained by the alternation of day and night: During sleeping hours there are few devices that consume power. However, during day-time hours there generally is increased activity, with the most electrical power being consumed in the evening hours. The low-rank component of the decomposition in Figure 2.4 captures the repeated general pattern. Meanwhile as intended, the group-sparse component identifies periods of larger electrical loads, caused by electrical devices that momentarily consumed power.

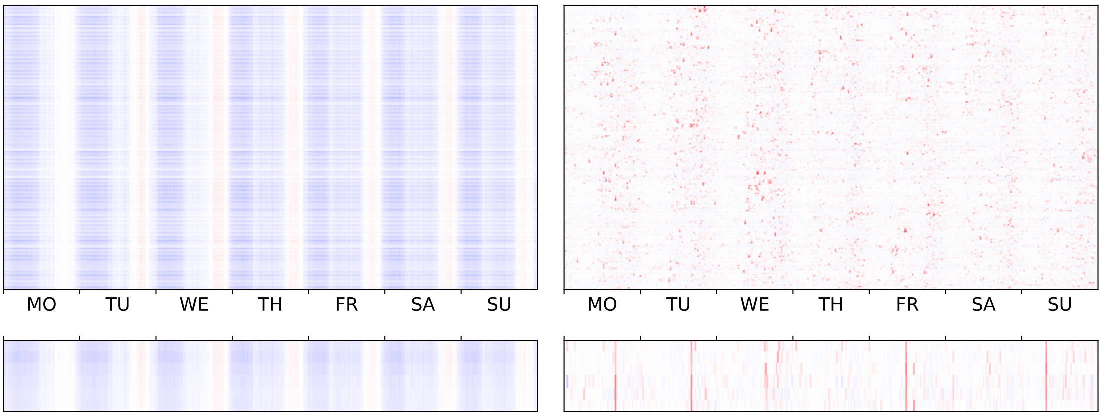

Figure 2.4: Decomposition of the electrical load profiles of 74 households over the course of one week. The left plot shows a typical repeated low-rank day-and-night pattern. The right plot shows the outlier component that captures periods of large loads, when some electrical devices consumed power. On the bottom, the decomposition for a single households is shown.

Cloud removal. Here, we investigate the task of detecting/removing clouds from satellite data. For this task, it makes sense to apply robust PCA because the surface does not change much (besides seasonal shifts in vegetation), while clouds cover parts of the surface only temporarily. We perform our experiments on a multi-spectral image time series that consists of 20 observations of Fort Wayne (Indiana, USA) from the years 2019 and 2020. The data was obtained from the *Copernicus Open Access Hub* [ESA, 2020]. After cropping and downsampling, each image has a size of 1000×1000 pixels and uses four bands: red, green, blue, and near infrared (these correspond to the bands 2, 3, 4, and 8 from the 13 available bands of the Sentinel-2 mission). To apply multi-view robust PCA, we group the four channels such that each pixel forms a group. In total, the data matrix has dimensions $\boldsymbol{X} \in \mathbb{R}^{4\,000\,000 \times 20}$. The results for $\gamma = 10^{-3}$ are shown in Figure 2.5. The outlier components capture the clouds such that the low-rank components are cloud-free images.

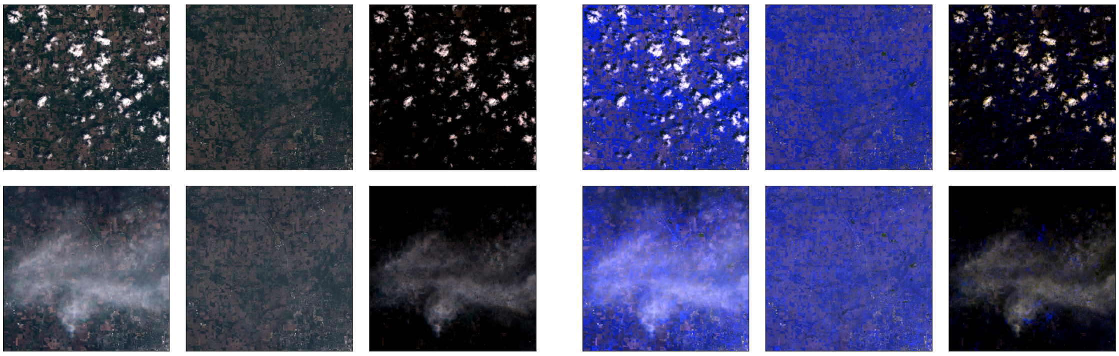

Figure 2.5: Robust PCA for a multi-spectral image time series (Sentinel-2 data). From left to right, the original images, the low-rank components, and the group-sparse components are shown. The outlier components separate the clouds from the surface.

Reconstruction of RGB images (multi-view data). Here, we briefly show that robust PCA for generalized multi-view models can be used to improve RGB images. We work with a multi-view dataset that consists of images from the *Amsterdam Library of Object Images* [Geusebroek et al., 2005], which is equipped with additional views from [Schubert and Zimek, 2019]. In the dataset, the data points are RGB images of the same object under 36 different light conditions. Each image has 144 × 192 pixels, where each pixel constitutes a different view of the image. Apart from that for each image, the first additional view consists of the first 13 Haralick features (radius 1 pixel), see [Haralick, 1979], and the second additional view is a standard RGB color histogram with 8 uniform bins. The whole data matrix has dimensions 82 965 × 36.

Exemplary results of applying multi-view robust PCA with $\gamma = 10^{-2}$ for two typical objects of the Amsterdam Library of Object Images are shown in Figure 2.6. In the low-rank component, spotlights and shadows have been reduced.

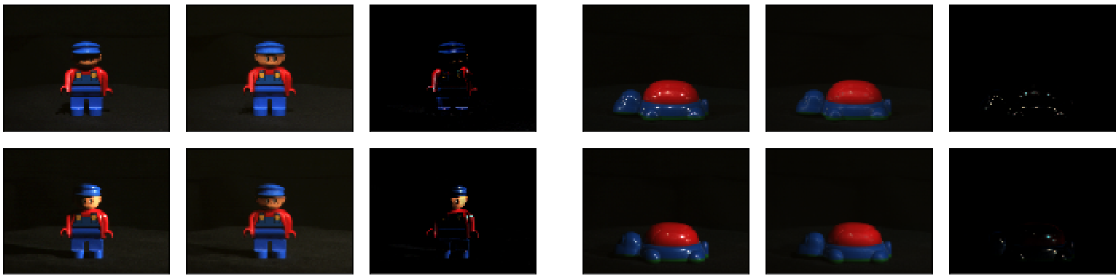

Figure 2.6: Robust recovery of RGB images for two objects from the Amsterdam Library of Object Images. From left to right, respectively, the original images, the reconstructed low-rank RGB images, and the outlier components are shown. Overexposures and shadows have been removed from the original images in the low-rank component and appear in the outlier component.

Detection of weather anomalies. The wave hindcast dataset *coastdat1*, which was obtained from [Helmholtz Centre for Materials and Coastal Research, 2012], contains a time series of wave conditions in the southern North Sea. The data that we use covers the year 2007 with a resolution of one hour. The covered area is $51.0N$ to $56.5N$ and $-3.0W$ to $10.5E$, using a grid size of approximately 0.05 degrees latitude and 0.10 degrees longitude. At each grid point, the sea state is described by the variables *significant weight height* (*hs*) and *mean wave period* (*mp*), which are derived from 2D wave spectra [Groll and Weisse, 2016].

The sea state at each grid position naturally defines a group of two parameters. Hence, to apply multi-view robust PCA for these groups, we change the data representation for a single time step from grid to a vector that contains the groups from all 6 324 sea-side grid positions. Hence, the data matrix has overall size $12\,648 \times 8\,760$, where each column corresponds to the data of one hour of the year.

The resulting decomposition for $\gamma = 10^{-3}$ can be found in Figure 2.7. Here, we only show the decomposition for selected time steps, and instead of the columns of the data matrix we directly show the covered area for the *mp* feature. We picked November, 9th as a special date since at this time there was a cyclone called Tilo that caused severe floods, that is, a strong weather anomaly. This is reflected in the outlier component in Figure 2.7, which highlights areas, where the storm was particularly strong. This experiment shows that generalized RPCA models with structured observations can also be used to detect anomalies.

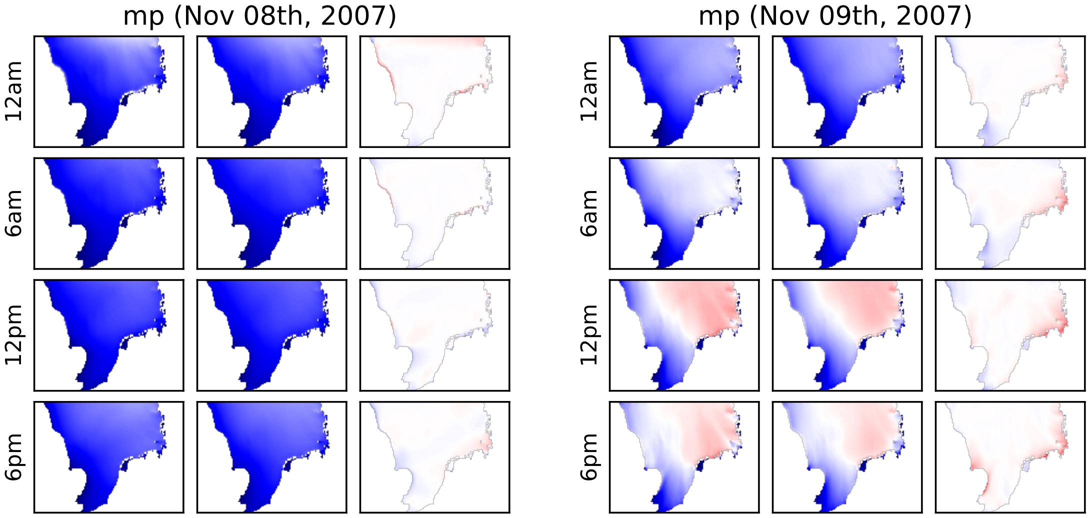

Figure 2.7: Wave hindcast data: The *mp* feature is shown from four time steps of November 9th, 2007 when cyclone Tilo caused severe North Sea floods (storm surges). From left to right, the columns show the original data, the low-rank components, and the outlier components. In the outlier components, the coastal lines show increased energy (red).

2.5 Concluding Remarks

In this chapter, we introduced robust principal component analysis for generalized multi-view models, where observations are structured in groups of measurements. A theoretically well-founded convex optimization problem can be used to separate principal components from groups of outliers. We empirically evaluated the rates of successful recovery for different decompositions using synthetic data. We presented a variety of real-world applications with naturally arising groups. The learned decompositions yield insights into the data, such as, general patterns and anomalies.

Future directions. The low-rank and group-sparse matrix decompositions from this chapter can be further generalized. In the introduction, we already mentioned general group structures that concern sub-matrices. Such group structures can stem from a data corruption mechanism, where sensor failures persist for several time steps in a time series of observations. As noted before, not much additional effort has to be placed in the theoretical analysis in order to show that exact recovery remains also possible when sub-matrices of the data matrix are corrupted. Next, we briefly introduce another possible extension of the model.

Groups with weights. In the generalized RPCA model of this chapter, we assumed data corruption caused by sensor failures that affects groups of measurements. In a scenario, where one has prior beliefs about sensor functionality, it may be useful to encode these beliefs in the learning problem. This can be done by using a weighted $\ell_{1,2}$-group norm

$$\|\boldsymbol{S}\|_{1,2}^{\boldsymbol{W}} = \sum_{i,j} w_{ij} \|\boldsymbol{s}_{ij}\|_2,$$

where $\boldsymbol{W} = (w_{ij})_{i\in[d],j\in[n]}$ is the matrix of weights. Here, a large weight w_{ij} means a low prior belief that the i-th sensor is malfunctioning for the j-th observation. If the prior belief of sensor functionality does not change over time, then $w_{ij} = w_i$ for all $j \in [n]$. A robust PCA model with prior beliefs can be learned by solving the problem

$$\min_{\boldsymbol{L},\,\boldsymbol{S}\,\in\mathbb{R}^{m\times n}} \|\boldsymbol{L}\|_* + \gamma\|\boldsymbol{S}\|_{1,2}^{\boldsymbol{W}} \quad \text{subject to} \quad \boldsymbol{X} = \boldsymbol{L} + \boldsymbol{S}, \tag{2.6}$$

where $\gamma > 0$ as usual. Another reason to use a weighted group norm may be that if all entries are of the same magnitude (which is the case after standardization of the data), the groups may be unbalanced in the sense that on average they have different norms. To balance out the effect of different group sizes, the (i, j)-th group could be weighted with $w_{ij} = w_i = 1/\sqrt{m_i}$. Similar weights have been used in [Lee and Hastie, 2015]. More generally, any non-negative weights can be used in Problem (2.6). It is possible to obtain similar theoretical guarantees for Problem (2.6) as for Problem (2.2) by following the same proof scheme. We briefly outline the changes to the proof in Appendix B.5.

It is not clear whether learning RPCA models using Problem (2.6) can improve upon results when learning with the basic Problem (2.2). Designing experiments to address this question could be the subject of future research.

Model evaluation. Finally, in Section 2.4 of this chapter, we demonstrated that the generalized RPCA model has many potential applications. However, the experiments on real-world data have been of a qualitative nature so far. Hence, it would be interesting to also quantitatively evaluate the learned RPCA models, that is, beyond visual inspection. For that, of course, suitable metrics would need to be defined. An example metric can be the *peak signal-to-noise ratio*, which has been used for image recovery in Lu et al. [2019].

Chapter 3

Fused Latent and Graphical Models

3.1 Introduction

In this chapter, we consider multivariate probability distributions with observed continuous and discrete variables that model pairwise interactions. Typically, not all the variables interact with each other. Hence, often one is interested in learning a sparse graphical model [Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010, 2011; Jalali et al., 2011; Lee and Hastie, 2015]. However, learning sparse graphical models may fail in the presence of latent variables because they induce indirect interactions of the observed variables. In the case of only a few latent quantitative variables, these interactions contribute an additional low-rank component to the interaction parameters of the marginal model for the observed variables. Hence, the pairwise interaction parameter matrix of the marginal model is characterized by a decomposition into a (group-)sparse component of direct interactions and a low-rank component of indirect interactions. Following [Chen et al., 2018], we call the resulting models *fused latent and graphical models*.

In the sequel of this introductory section, we approach fused latent and graphical models from two different perspectives: The first perspective in Section 3.1.1 is based on the work [Nussbaum and Giesen, 2020b]. It relates fused latent and graphical models to *factor models*. In the course, we introduce a convex optimization problem for learning fused latent and graphical models. The second perspective in Section 3.1.2 is based on the work [Nussbaum and Giesen, 2019a]. It shows that the same optimization problem can be obtained as the dual of a maximum-entropy problem with a new type of relaxation, where the sample means collectively need to match the expected values only up to a given tolerance. Finally, in Section 3.1.3 we summarize fused latent and graphical models and give an overview of the remaining content of this chapter.

3.1.1 Factor models with direct interactions

In this section, we deduce fused latent and graphical models from factor models. As special cases we discuss Gaussian, discrete, and mixed-type distributions.

Gaussian models. *Multivariate Gaussians* $p(\boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ are still among the most popular probabilistic models on multidimensional quantitative sample spaces $\mathcal{Y} = \mathbb{R}^q$. A single model has q parameters for the mean vector $\boldsymbol{\mu}$ and $\binom{q}{2} + q$ pairwise parameters for the symmetric covariance matrix $\boldsymbol{\Sigma}$. Hence, since the number of parameters is fairly large, estimating them is prone to overfitting. Moreover, the maximum likelihood estimate of the covariance matrix given by the empirical covariance matrix is not regular if there are less than q data points. *Factor analysis* that was developed by Spearman [1904] while working on a theory of human abilities, is used to address these shortcomings. *Factor models* assume a small number of unobserved (latent) variables called factors. These factors describe all correlations among the observed variables.

The sample space $\mathcal{Y} \times \mathcal{Z} = \mathbb{R}^q \times \mathbb{R}^r$ of a Gaussian factor model decomposes into an observed part $\mathcal{Y} = \mathbb{R}^q$ and an unobserved part $\mathcal{Z} = \mathbb{R}^r$, typically with r much smaller than q. The Gaussian factor model is a multivariate Gaussian on $\mathcal{Y} \times \mathcal{Z}$ with marginals $p(\boldsymbol{y}) \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Psi} + \boldsymbol{\Gamma}\boldsymbol{\Gamma}^\mathsf{T})$ and $p(\boldsymbol{z}) \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I})$, where $\boldsymbol{\Gamma} \in \mathbb{R}^{q\times r}$, $\boldsymbol{\Psi} \in \mathbb{R}^{q\times q}$ is a full-rank diagonal matrix, and $\boldsymbol{I} \in \mathbb{R}^{r\times r}$ is the identity matrix. Thus, the covariance matrix of the marginal distribution $p(\boldsymbol{y})$ for the observed variables is the sum of the diagonal (sparse) matrix $\boldsymbol{\Psi}$ and the low-rank matrix $\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\mathsf{T}$, where $\boldsymbol{\Gamma}$ describes the correlation of the observed variables with the unobserved variables. Typically, the parameters $\boldsymbol{\mu}$, $\boldsymbol{\Gamma}$, and $\boldsymbol{\Psi}$ of the Gaussian factor model are estimated using the *expectation maximization (EM)* algorithm, see [Dempster et al., 1977].

Chandrasekaran et al. [2012] introduced an alternative to the Gaussian factor model that also addresses the shortcomings of multivariate Gaussians. As in the standard Gaussian factor model, they assume that the joint distribution $p(\boldsymbol{y}, \boldsymbol{z})$ of the observed and unobserved variables is a multivariate Gaussian. The *precision matrix* (inverse of the covariance matrix) of the marginal distribution $p(\boldsymbol{y})$ for the observed variables can be obtained from the precision matrix of the joint distribution $\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_y & \boldsymbol{\Lambda}_{yz} \\ \boldsymbol{\Lambda}_{zy} & \boldsymbol{\Lambda}_z \end{pmatrix}$ as the upper Schur complement $\boldsymbol{\Lambda}_y - \boldsymbol{\Lambda}_{yz}\boldsymbol{\Lambda}_z^{-1}\boldsymbol{\Lambda}_{zy}$. Thus, the precision matrix of the marginal distribution $p(\boldsymbol{y})$ is the sum of the matrix $\boldsymbol{\Lambda}_y$, which Chandrasekaran et al. [2012] assume to be sparse, and the low-rank matrix $-\boldsymbol{\Lambda}_{yz}\boldsymbol{\Lambda}_z^{-1}\boldsymbol{\Lambda}_{zy}$. Note that while entries in the covariance matrix measure correlation, they measure conditional dependence in the precision matrix. Still, we refer to a low-rank approximation of either matrix as a factor model. Chandrasekaran et al. [2012] show that the parameters $\boldsymbol{S} = \boldsymbol{\Lambda}_y$ and $\boldsymbol{L} = \boldsymbol{\Lambda}_{yz}\boldsymbol{\Lambda}_z^{-1}\boldsymbol{\Lambda}_{zy}$ can be estimated consistently in the high-dimensional setting, where the dimensions q and r are allowed to grow with the number of sample points. They learn the parameters through the convex optimization problem

$$\min_{\boldsymbol{S},\, \boldsymbol{L} \in \mathrm{Sym}(q)} \ell(\boldsymbol{S}-\boldsymbol{L}) + \lambda\big(\gamma\|\boldsymbol{S}\|_1 + \mathrm{tr}(\boldsymbol{L})\big) \quad \text{subject to} \quad \boldsymbol{S}-\boldsymbol{L} \succ \boldsymbol{0},\, \boldsymbol{L} \succeq \boldsymbol{0},$$

where $\mathrm{Sym}(q)$ is the set of symmetric $(q \times q)$-matrices, and

$$\ell(\boldsymbol{S}-\boldsymbol{L}) = \langle \boldsymbol{S}-\boldsymbol{L}, \hat{\boldsymbol{\Sigma}} \rangle - \log\det(\boldsymbol{S}-\boldsymbol{L})$$

is the negative log-likelihood of a zero-mean multivariate Gaussian distribution that uses the standard inner product $\langle \cdot, \cdot \rangle$ for matrices. Here, the empirical second-moment matrix $\hat{\boldsymbol{\Sigma}} = 1/n \sum_{k=1}^{n} \boldsymbol{y}^{(k)}[\boldsymbol{y}^{(k)}]^{\mathsf{T}}$ has been computed from observed data points $\boldsymbol{y}^{(1)}, \ldots, \boldsymbol{y}^{(n)} \in \mathcal{Y}$. Moreover, as before, the ℓ_1-norm regularization term $\|\boldsymbol{S}\|_1$ induces sparsity on $\boldsymbol{S}$, and the trace/nuclear norm regularization term $\mathrm{tr}(\boldsymbol{L})$ induces low rank for positive semidefinite $\boldsymbol{L} \succeq \boldsymbol{0}$. Note that for $\boldsymbol{L} \succeq \boldsymbol{0}$ it holds $\|\boldsymbol{L}\|_* = \mathrm{tr}(\boldsymbol{L})$. Finally, $\lambda > 0$ and $\gamma > 0$ are regularization parameters that provide trade-offs between the different parts of the objective function. Specifically, the parameter λ controls the influence of the negative log-likelihood term in relation to the regularization terms. The second trade-off parameter γ determines the relative weights of the regularization terms.

It is worth noting that the standard Gaussian factor model induces a sparse + low-rank structure on the precision matrix as well. The joint probability density function for the observed and unobserved variables of the Gaussian factor model is given as

$$p(\boldsymbol{y}, \boldsymbol{z}) \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu} \\ \boldsymbol{0} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Psi} + \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathsf{T}} & \boldsymbol{\Gamma} \\ \boldsymbol{\Gamma}^{\mathsf{T}} & \boldsymbol{I} \end{pmatrix}\right).$$

It can easily be checked by inverting the covariance matrix that the precision matrix of the joint model is given by $\begin{pmatrix} \boldsymbol{\Psi}^{-1} & -\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma} \\ -\boldsymbol{\Gamma}^{\mathsf{T}}\boldsymbol{\Psi}^{-1} & \boldsymbol{I} + \boldsymbol{\Gamma}^{\mathsf{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma} \end{pmatrix}$. Therefore, the precision matrix of the marginal distribution, again given by the upper Schur complement, reads as $\boldsymbol{\Psi}^{-1} - \boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}(\boldsymbol{I} + \boldsymbol{\Gamma}^{\mathsf{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma})\boldsymbol{\Gamma}^{\mathsf{T}}\boldsymbol{\Psi}^{-1}$. Note that $\boldsymbol{S} = \boldsymbol{\Psi}^{-1}$ is a sparse diagonal matrix that does not permit direct interactions among the observed variables. Moreover, the matrix $\boldsymbol{L} = \boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma}(\boldsymbol{I} + \boldsymbol{\Gamma}^{\mathsf{T}}\boldsymbol{\Psi}^{-1}\boldsymbol{\Gamma})\boldsymbol{\Gamma}^{\mathsf{T}}\boldsymbol{\Psi}^{-1}$ has at most rank r. Thus, the sparse + low-rank decomposition in the fused latent and graphical model that Chandrasekaran et al. [2012] proposed is more flexible than the standard Gaussian factor model since it allows for direct interactions among the observed variables.

Binary and discrete models. Factor models have also been discussed for other distributions, among them Ising models on the sample space $\mathcal{X} = \{0, 1\}^d$. Ising models are specified by a symmetric pairwise interaction matrix $\boldsymbol{Q} \in \mathrm{Sym}(d)$, that is, $p(\boldsymbol{x}) \propto \exp\big(1/2\, \boldsymbol{x}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{x}\big)$ for $\boldsymbol{x} \in \mathcal{X}$. As in the Gaussian case, the number of parameters in the interaction matrix $\boldsymbol{Q}$ can be too large to be estimated faithfully from a small data sample. A common approach for dealing with this is to assume that $\boldsymbol{Q}$ is sparse, which induces a sparse graphical model structure.

While a sparse graphical model structure is often a good choice when modeling phenomena in physics, Marsman et al. [2015] emphasized that a low-rank structure serves many social science applications better. Low-rank structures are usually induced by unobserved latent traits such as, for example, the intelligence or extroversion of test takers. In fact, many psychometric tests are modeled using item response theory (IRT), see [Embretson and Reise, 2013]. In [Marsman et al., 2018] it is shown that IRT models are intimately related to Ising models. While classical IRT only considers the dichotomized outcomes *right* and *wrong* for each question, *polytomous* IRT, see Ostini and Nering [2006], allows more general discrete outcomes. Apart from right and wrong there can, for instance, be an additional *no-choice* option. Alternatively, all available options from multiple-choice questions can be taken into account. Hence, in general, IRT considers models with observed and unobserved parts. The observed part consists of variables $\boldsymbol{x}$ from a *discrete* sample space $\mathcal{X} = \prod_{i=1}^{d} \mathcal{X}_i$, where the $\mathcal{X}_i = \{0, \dots, m_i\}$ are finite sets of choice options. The unobserved part is composed of additional latent variables $\boldsymbol{z}$ from a *continuous* sample space $\mathcal{Z} = \mathbb{R}^r$. IRT models can be seen as low-rank factor models due to the usually small number of latent variables. On the sample space $\mathcal{X} \times \mathcal{Z} = \prod_{i=1}^{d} \mathcal{X}_i \times \mathbb{R}^r$, we consider a joint probability distribution that is a pairwise *conditional Gaussian (CG)* distribution, see [Lauritzen, 1996]. It is defined as

$$p(\boldsymbol{x}, \boldsymbol{z}) \propto \exp\left(\frac{1}{2}\,\overline{\boldsymbol{x}}^{\mathsf{T}}\boldsymbol{Q}\,\overline{\boldsymbol{x}} + \boldsymbol{z}^{\mathsf{T}}\boldsymbol{R}\,\overline{\boldsymbol{x}} - \frac{1}{2}\,\boldsymbol{z}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{z}\right), \qquad (\boldsymbol{x}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{Z}, \tag{3.1}$$

where $\boldsymbol{Q} \in \mathrm{Sym}(m)$ with $m = \sum_{i=1}^{d} m_i$ describes interactions among the observed discrete variables, $\boldsymbol{R} \in \mathbb{R}^{r \times m}$ describes interactions between the observed discrete and unobserved quantitative variables, and $\boldsymbol{0} \prec \boldsymbol{\Lambda} \in \mathrm{Sym}(r)$ describes interactions among the unobserved quantitative variables. Moreover, for $\boldsymbol{x} \in \mathcal{X}$, we define the concatenated indicator variables as

$$\begin{aligned} &\overline{\boldsymbol{x}} = (\overline{\boldsymbol{x}}_1, \dots, \overline{\boldsymbol{x}}_d) \in \{0,1\}^m, \quad \text{where} \\ &\overline{\boldsymbol{x}}_i = (\mathbb{1}[x_i = 1], \dots, \mathbb{1}[x_i = m_i]) \in \{0,1\}^{m_i}. \end{aligned} \tag{3.2}$$

Here, we left out the indicator variables for the state zero, respectively. This is to ensure a unique parametrization of Model (3.1). The conditional densities $p(\boldsymbol{z} \,|\, \boldsymbol{x})$ in Model (3.1) are r-variate Gaussians on $\mathcal{Z}$, hence the name CG (conditional Gaussian) distribution. Now, the marginal distribution of the discrete variables in $\mathcal{X}$ is obtained by integrating over the unobserved variables in $\mathcal{Z}$ (see Appendix C.2). It is given as

$$p(\boldsymbol{x}) \propto \exp\left(\frac{1}{2}\,\overline{\boldsymbol{x}}^{\mathsf{T}}\Big(\boldsymbol{Q} + \boldsymbol{R}^{\mathsf{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{R}\Big)\,\overline{\boldsymbol{x}}\right), \qquad \boldsymbol{x} \in \mathcal{X}.$$

The matrix $\boldsymbol{L} = \boldsymbol{R}^{\mathsf{T}}\boldsymbol{\Lambda}^{-1}\boldsymbol{R}$ is symmetric and positive semidefinite. Hence, in the marginal model the interaction matrix is the sum $\boldsymbol{Q} + \boldsymbol{L}$, where $\boldsymbol{Q}$ describes direct interactions among the observed variables and $\boldsymbol{L}$ describes additional indirect interactions induced by the unobserved variables. For a purely low-rank model one assumes that $\boldsymbol{Q}$ is diagonal, prohibiting direct interactions. In [Nussbaum and Giesen, 2019a] (binary variables) and [Nussbaum and Giesen, 2020a] (general discrete variables), we considered fused latent and graphical models, where $\boldsymbol{S} = \boldsymbol{Q}$ is (group) sparse. Here, the groups are given by

$$\boldsymbol{S} = \left(\begin{array}{c|c|c|c} \boldsymbol{S}_{11} & \boldsymbol{S}_{12} & \cdots & \boldsymbol{S}_{1d} \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \boldsymbol{S}_{d1} & \boldsymbol{S}_{d2} & \cdots & \boldsymbol{S}_{dd} \end{array}\right) \in \mathrm{Sym}(m),$$

where for $i, j \in [d]$ the group $\boldsymbol{S}_{ij} \in \mathbb{R}^{m_i \times m_j}$ contains the parameters that describe the interaction between the i-th and j-th observed variable. Similarly as for the Gaussian case [Chandrasekaran et al., 2012], discrete fused latent and graphical models can be learned using the following convex optimization problem:

$$\min_{\boldsymbol{S},\, \boldsymbol{L} \in \mathrm{Sym}(m)} \quad \ell(\boldsymbol{S} + \boldsymbol{L}) + \lambda\left(\gamma \|\boldsymbol{S}\|_{1,2} + \mathrm{tr}(\boldsymbol{L})\right) \quad \text{subject to} \quad \boldsymbol{L} \succeq \boldsymbol{0}. \tag{3.3}$$

Here, again $\lambda, \gamma > 0$ are trade-off parameters, $\|\boldsymbol{S}\|_{1,2} = \sum_{i,j\in[d]} \|\boldsymbol{S}_{ij}\|_2$ is the $\ell_{1,2}$-norm that specializes to the ℓ_1-norm if all observed variables are binary, and as we show in Appendix C.2,

$$\ell(\boldsymbol{S} + \boldsymbol{L}) = 2a(\boldsymbol{S} + \boldsymbol{L}) - \langle \boldsymbol{S} + \boldsymbol{L}, \hat{\boldsymbol{\Sigma}} \rangle$$

is the (rescaled) negative log-likelihood for the model $p(\boldsymbol{x}) = \exp(1/2\, \overline{\boldsymbol{x}}^\mathsf{T}(\boldsymbol{S} + \boldsymbol{L})\, \overline{\boldsymbol{x}} - a(\boldsymbol{S} + \boldsymbol{L}))$ with log-partition (normalization) function a and the empirical second-moment matrix $\hat{\boldsymbol{\Sigma}} = 1/n \sum_{k=1}^{n} \overline{\boldsymbol{x}}^{(k)}\, [\overline{\boldsymbol{x}}^{(k)}]^\mathsf{T}$, which has been computed from n indicator-encoded observations $\overline{\boldsymbol{x}}^{(1)}, \ldots, \overline{\boldsymbol{x}}^{(n)} \in \{0,1\}^m$.

Mixed models. In this thesis, we consider a general pairwise model that can account for observed discrete variables in $\mathcal{X} = \prod_{i=1}^{d} \mathcal{X}_i = \prod_{i=1}^{d} \{0, \ldots, m_i\}$ as well as for observed quantitative variables in $\mathcal{Y} = \mathbb{R}^q$. Factor models have also been investigated in this setting, see, for example, [Bartholomew and Knott, 1999]. Particularly, they are special cases of general exponential family factor models with mixed observed variables that have been considered in [Sammel et al., 1997] and [Wedel and Kamakura, 2001]. As above, the low-rank structure can be obtained by marginalizing out the latent variables in $\mathcal{Z} = \mathbb{R}^r$ from the full model on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z} = \prod_{i=1}^{d} \{0, \ldots, m_i\} \times \mathbb{R}^q \times \mathbb{R}^r$ with distribution

$$p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \propto \exp\left(\frac{1}{2}\, \overline{\boldsymbol{x}}^\mathsf{T} \boldsymbol{Q}\, \overline{\boldsymbol{x}} + (\boldsymbol{y}, \boldsymbol{z})^\mathsf{T} \boldsymbol{R}\, \overline{\boldsymbol{x}} - \frac{1}{2} (\boldsymbol{y}, \boldsymbol{z})^\mathsf{T} \boldsymbol{\Lambda} (\boldsymbol{y}, \boldsymbol{z})\right),$$

where now the interaction parameter matrices $\boldsymbol{\Lambda}$ and $\boldsymbol{R}$ respectively decompose into interactions with the observed and with the latent quantitative variables, that is, they are structured as $\boldsymbol{R} = \begin{pmatrix} \boldsymbol{R}_y \\ \boldsymbol{R}_z \end{pmatrix} \in \mathbb{R}^{(q+r)\times m}$ and $\boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_y & \boldsymbol{\Lambda}_{yz} \\ \boldsymbol{\Lambda}_{zy} & \boldsymbol{\Lambda}_z \end{pmatrix} \in \mathrm{Sym}(q+r)$. The marginal distribution on $\mathcal{X} \times \mathcal{Y}$, see again Appendix C.2, is given by the fused latent and graphical model

$$p(\boldsymbol{x}, \boldsymbol{y}) \propto \exp\left(\frac{1}{2}\,(\overline{\boldsymbol{x}}, \boldsymbol{y})^\mathsf{T}\,(\boldsymbol{S} + \boldsymbol{L})\,(\overline{\boldsymbol{x}}, \boldsymbol{y})\right), \quad (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y},$$

where $\boldsymbol{S} = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{R}_y^\mathsf{T} \\ \boldsymbol{R}_y & \text{-}\boldsymbol{\Lambda}_y \end{pmatrix}$ and $\boldsymbol{L} = \begin{pmatrix} \boldsymbol{R}_z & \text{-}\boldsymbol{\Lambda}_{zy} \end{pmatrix}^\mathsf{T} \boldsymbol{\Lambda}_z^{-1} \begin{pmatrix} \boldsymbol{R}_z & \text{-}\boldsymbol{\Lambda}_{zy} \end{pmatrix} \succeq \boldsymbol{0}$. Denote by $\Lambda[\boldsymbol{S} + \boldsymbol{L}]$ the interaction parameters between the quantitative variables in the marginal model. Then, the model above is only normalizable if $\Lambda[\boldsymbol{S} + \boldsymbol{L}] \succ \boldsymbol{0}$. Note that technically, for an interaction parameter matrix of the form $\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{R}^\mathsf{T} \\ \boldsymbol{R} & \text{-}\boldsymbol{\Lambda} \end{pmatrix} \in \mathbb{R}^{(m+q)\times(m+q)}$, we let $\Lambda[\boldsymbol{\Theta}] = \boldsymbol{\Lambda}$, that is, $\Lambda[\boldsymbol{S} + \boldsymbol{L}]$ extracts the interaction parameters between the quantitative variables from the bottom right block of the parameter matrix.

When $\boldsymbol{S}$ is group sparse in the fused latent and graphical model above, then this model can be learned using the familiar problem

$$\min_{\boldsymbol{S}, \boldsymbol{L} \in \mathrm{Sym}(m+q)} \quad \ell(\boldsymbol{S} + \boldsymbol{L}) + \lambda\,(\gamma \|\boldsymbol{S}\|_{1,2} + \mathrm{tr}(\boldsymbol{L})) \quad \text{s.t.} \quad \Lambda[\boldsymbol{S} + \boldsymbol{L}] \succ \boldsymbol{0},\, \boldsymbol{L} \succeq \boldsymbol{0}, \tag{3.4}$$

where again $\lambda, \gamma > 0$ are trade-off parameters and

$$\ell(\boldsymbol{S} + \boldsymbol{L}) = 2a(\boldsymbol{S} + \boldsymbol{L}) - \langle \boldsymbol{S} + \boldsymbol{L}, \hat{\boldsymbol{\Sigma}} \rangle$$

is the (rescaled) negative log-likelihood with log partition-function a and empirical second-moment matrix $\hat{\boldsymbol{\Sigma}}$ of the observed (indicator-encoded) discrete and quantitative variables, see again Appendix C.2. Observe that Problem (3.4) generalizes the respective problems for learning discrete and Gaussian fused latent and graphical models that we have discussed earlier.

3.1.2 A new relaxation of the maximum-entropy principle

Here, we provide another perspective on fused latent and graphical models that is based on the principle of maximum entropy, which was proposed by Jaynes [1957] for probability density estimation. It states that from the probability densities that represent the current state of knowledge one should choose the one with the largest entropy, that is, the one which does not introduce additional biases. The state of knowledge is often given by sample points from a sample space and some fixed functions (sufficient statistics) on the sample space. The knowledge is then encoded naturally in form of constraints on the probability density by requiring that the expected values of the functions equal their respective sample means.

Here, we consider discrete distributions only because entropy is something that we believe is inherently discrete (although generalizations to non-discrete variables have been attempted, see Appendix C.1.2 for a short discussion). The binary case was covered in [Nussbaum and Giesen, 2019a]. Here, we consider the general discrete case that uses the multivariate sample space $\mathcal{X} = \prod_{i=1}^{d} \mathcal{X}_i$ with $\mathcal{X}_i = \{0, \dots, m_i\}$ and functions

$$f_{ij;kl} : \mathcal{X} \mapsto \{0,1\}, \boldsymbol{x} \mapsto \mathbb{1}[x_i = k, x_j = l] \quad \text{for } i,j \in [d], k \in [m_i], l \in [m_j].$$

Suppose we are given sample points $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)} \in \mathcal{X}$. Then formally, for estimating the distribution from which the sample points are drawn, the principle of maximum entropy suggests solving the following entropy maximization problem

$$\max_{p \in \mathcal{P}} H(p) \quad \text{s.t.} \quad \mathbb{E}[f_{ij;kl}] = \frac{1}{n} \sum_{k=1}^{n} f_{ij;kl}(\boldsymbol{x}^{(k)}) \text{ for all } i,j \in [d], k \in [m_i], l \in [m_j],$$

where $\mathcal{P}$ is the set of all probability distributions on $\mathcal{X}$, the expectation is with respect to the distribution $p \in \mathcal{P}$, and $H(p) = -\mathbb{E}[\log p]$ is the entropy. We denote the matrix of functions $\left(f_{ij;kl}\right)_{k\in[m_i],\, l\in[m_j]} : \mathcal{X} \to \{0,1\}^{m_i \times m_j}$ that correspond to the variables i and j by Σ_{ij}. We summarize all functions as

$$\Sigma : \mathcal{X} \to \{0,1\}^{m\times m},\, \boldsymbol{x} \mapsto \left(\begin{array}{c|c|c|c} \Sigma_{11}(\boldsymbol{x}) & \Sigma_{12}(\boldsymbol{x}) & \cdots & \Sigma_{1d}(\boldsymbol{x}) \\ \hline \vdots & \vdots & \ddots & \vdots \\ \hline \Sigma_{d1}(\boldsymbol{x}) & \Sigma_{d2}(\boldsymbol{x}) & \cdots & \Sigma_{dd}(\boldsymbol{x}) \end{array}\right),$$

where as before $m = \sum_{i=1}^{d} m_i$. Similarly, we use the compact notation

$$\hat{\boldsymbol{\Sigma}} = \left(\frac{1}{n} \sum_{a=1}^{n} f_{ij;kl}(\boldsymbol{x}^{(a)})\right)_{i,j\in[d],\, k\in[m_i],\, l\in[m_j]}$$

for the $(m \times m)$-matrix of sample means. Now, the entropy maximization problem becomes

$$\max_{p \in \mathcal{P}} H(p) \quad \text{s.t. } \mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}} = \boldsymbol{0}.$$

Dudík et al. [2004] observed that invoking the principle of maximum entropy tends to overfit when the number of features is large. Requiring that the expected values of the functions equal their respective sample means can be too restrictive. Consequently, in the binary case, where $m_i = 1$ for all i, they proposed to relax the constraint using the maximum norm as

$$\|\mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}}\|_\infty \leq c$$

for some $c > 0$. The relaxation implies that for every function, the expected value only needs to match the sample mean up to a tolerance of c. In the general discrete case, instead of using the ℓ_∞-norm, it makes more sense to relax the constraint as

$$\|\mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}}\|_{\infty,2} \leq c,$$

where the $\ell_{\infty,2}$-norm is defined as $\|\boldsymbol{A}\|_{\infty,2} = \max_{i,j\in[d]} \|\boldsymbol{A}_{ij}\|_2$ for a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times m}$ with groups $\boldsymbol{A}_{ij} \in \mathbb{R}^{m_i\times m_j}$. With this relaxation, for each pair of variables, the associated functions that describe an interaction between these variables are grouped. Thus, the parameter c specifies how much the expected values are allowed to deviate from their corresponding sample means, separately for each group in the ℓ_2-norm sense.

The dual of the relaxed problem has a natural interpretation as a group-selective $\ell_{1,2}$-regularized log-likelihood maximization problem

$$\max_{\boldsymbol{S}\,\in\,\mathrm{Sym}(m)} \ \ell(\boldsymbol{S}) - c\|\boldsymbol{S}\|_{1,2}.$$

Here,

$$\ell(\boldsymbol{S}) = \langle \boldsymbol{S}, \hat{\boldsymbol{\Sigma}}\rangle - a(\boldsymbol{S}) \tag{3.5}$$

is the log-likelihood function for the discrete model $p(\boldsymbol{x}) = \exp\big(\overline{\boldsymbol{x}}^\top \boldsymbol{S}\,\overline{\boldsymbol{x}} - a(\boldsymbol{S})\big)$ (note that here we omitted the factor 1/2 that we elsewhere include in the model definition). The normalizer (log-partition function) is $a(\boldsymbol{S}) = \log(\sum_{\boldsymbol{x}\in\mathcal{X}} \langle \boldsymbol{S}, \Sigma(\boldsymbol{x})\rangle)$.

The key to our model is a restriction of the relaxation of the entropy maximization problem, which is obtained by also enforcing the alternative constraint

$$\|\mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}}\| \leq \lambda,$$

where $\lambda > 0$ and $\|\cdot\|$ denotes the spectral norm. A difference to the $\ell_{\infty,2}$-norm constraint is that now the expected values of *all* functions only need to collectively match their sample means up to a tolerance of λ, instead of group-wise. The dual of the more strictly relaxed entropy maximization problem

$$\max_{p\,\in\,\mathcal{P}} H(p) \quad \text{s.t.} \quad \|\mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}}\|_{\infty,2} \leq c \ \text{ and } \ \|\mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}}\| \leq \lambda$$

is the regularized log-likelihood maximization problem

$$\max_{\boldsymbol{S},\,\boldsymbol{L}\,\in\,\mathrm{Sym}(m)} \ \ell(\boldsymbol{S}+\boldsymbol{L}) - c\|\boldsymbol{S}\|_{1,2} - \lambda\|\boldsymbol{L}\|_*,$$

see Appendix C.1. Here, as before $\|\cdot\|_*$ is the nuclear norm, which promotes low rank on $\boldsymbol{L}$. Thus, a solution of the dual problem is the sum of a group-sparse matrix $\boldsymbol{S}$ and a low-rank matrix $\boldsymbol{L}$. Again, this can be interpreted as follows: The variables interact indirectly through the low-rank matrix $\boldsymbol{L}$, while some of the direct interactions through the matrix $\boldsymbol{S}$ are turned off by setting groups of entries in $\boldsymbol{S}$ to zero.

We get a more intuitive interpretation of the dual problem if we consider a weaker version of the spectral norm constraint. The spectral norm constraint is equivalent to the two constraints

$$\mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}} \preceq \lambda \boldsymbol{I} \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} - \mathbb{E}[\Sigma] \preceq \lambda \boldsymbol{I}$$

that bound the spectrum of the matrix $\mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}}$ from above and below. If we replace the spectral norm constraint by only the second of these two constraints in the maximum-entropy problem, then the dual problem becomes

$$\max_{\boldsymbol{S},\, \boldsymbol{L} \,\in\, \mathrm{Sym}(m)} \ell(\boldsymbol{S} + \boldsymbol{L}) - c\|\boldsymbol{S}\|_{1,2} - \lambda \operatorname{tr}(\boldsymbol{L}) \quad \text{s.t.} \quad \boldsymbol{L} \succeq \mathbf{0}.$$

This problem is equivalent to Problem (3.3) from Section 3.1.1.

3.1.3 Setting and outlook for this chapter

To summarize, we consider fused latent graphical models for pairwise conditional Gaussian distributions on the sample space $\mathcal{X} \times \mathcal{Y} = \prod_{i=1}^{d}\{0, \dots, m_i\} \times \mathbb{R}^q$. They are characterized by a decomposition of the pairwise interaction parameter matrix $\boldsymbol{\Theta}$ into a group-sparse component $\boldsymbol{S}$ of direct interactions and a low-rank component $\boldsymbol{L}$ of indirect interactions. Consequently, we consider fused and latent graphical models of the form

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}) &= \exp\left\{\frac{1}{2}(\overline{\boldsymbol{x}}, \boldsymbol{y})^{\mathsf{T}} \boldsymbol{\Theta} (\overline{\boldsymbol{x}}, \boldsymbol{y}) - a(\boldsymbol{\Theta})\right\} \\ &= \exp\left\{\frac{1}{2}(\overline{\boldsymbol{x}}, \boldsymbol{y})^{\mathsf{T}} (\boldsymbol{S} + \boldsymbol{L}) (\overline{\boldsymbol{x}}, \boldsymbol{y}) - a(\boldsymbol{S} + \boldsymbol{L})\right\}, \qquad (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}. \quad (3.6) \end{aligned}$$

Here, $\overline{\boldsymbol{x}}$ are the concatenated indicator variables, see (3.2), and a is the log-partition function. Using $w = m + q = \sum_{i=1}^{d} m_i + q$, the parameters are restricted to $\boldsymbol{S} \in \mathrm{Sym}(w)$ and $\mathbf{0} \preceq \boldsymbol{L} \in \mathrm{Sym}(w)$. Moreover, (3.6) only defines a valid model if the interaction parameters $\Lambda[\boldsymbol{S} + \boldsymbol{L}]$ between the quantitative variables satisfy the condition $\Lambda[\boldsymbol{S} + \boldsymbol{L}] \succ \mathbf{0}$ that is necessary for normalizability. Given n observations $(\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)}) \in \mathcal{X} \times \mathcal{Y}$, we estimate fused latent and graphical models using the convex problem

$$\min_{\boldsymbol{S},\, \boldsymbol{L} \,\in\, \mathrm{Sym}(w)} \ell(\boldsymbol{S} + \boldsymbol{L}) + \lambda\left(\gamma \|\boldsymbol{S}\|_{1,2} + \operatorname{tr}(\boldsymbol{L})\right) \quad \text{s.t.} \quad \Lambda[\boldsymbol{S} + \boldsymbol{L}] \succ \mathbf{0},\, \boldsymbol{L} \succeq \mathbf{0}. \quad (3.7)$$

Here, the (rescaled) negative log-likelihood is given by

$$\ell(\boldsymbol{S} + \boldsymbol{L}) = 2a(\boldsymbol{S} + \boldsymbol{L}) - \langle \boldsymbol{S} + \boldsymbol{L}, \hat{\boldsymbol{\Sigma}} \rangle,$$

where $\hat{\boldsymbol{\Sigma}} = 1/n \sum_{k=1}^{n} (\overline{\boldsymbol{x}}^{(k)}, \boldsymbol{y}^{(k)})(\overline{\boldsymbol{x}}^{(k)}, \boldsymbol{y}^{(k)})^{\mathsf{T}}$ is the empirical second-moment matrix.

In the remaining part of this chapter, we study Problem (3.7). For that, we proceed as follows: In Section 3.2, we show that fused latent and graphical models can be recovered consistently by solving instances of Problem (3.7). In the spirit of previous works on sparse graphical models [Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010, 2011; Lee and Hastie, 2015] and the seminal paper on fused latent and graphical models by Chandrasekaran et al. [2012], we consider the high-dimensional setting. In this setting, the number of observed variables and the number of latent quantitative variables are allowed to grow with the number of observed samples. In the seminal work [Chandrasekaran et al., 2012], consistency was shown for Gaussian variables. In [Nussbaum and Giesen, 2019a, 2020a,b], we generalized the consistency result to distributions with observed binary and quantitative variables. In this thesis, we consider the even more general case of distributions with observed *discrete* and quantitative variables. Thus, the consistency result that we formulate in Section 3.2 encompasses all previously examined cases as special cases.

After the theoretical analysis, we introduce a practical solver for learning general fused latent and graphical models in Section 3.3. Next, in Section 3.4, we support our theoretical findings with experiments on synthetic and real-world data from polytomous item response theory studies. Finally, we venture on an excursion to a variant of Benson's algorithm [Benson, 1998] that we develop for selecting the regularization parameters of Problem (3.7) in a principled and efficient way.

3.2 Consistency Results

In this section, we motivate and state our main result that can be summarized as follows: The convex optimization Problem (3.7) allows for consistent recovery of a (group-)sparse + low-rank decomposition of the pairwise interaction parameter matrix of a conditional Gaussian (CG) distribution. This consistency holds in the high-dimensional setting, where we permit the number of data points n, the number of observed discrete variables d, the number of observed quantitative variables q, and the number of latent variables r to grow simultaneously. At the same time, we assume the geometry of the problem, which is given by the curvature of the likelihood function, to be fixed. Now, before we state our result, we specify the framework of our consistency analysis in more detail and elaborate conditions that need to be satisfied for consistent recovery.

Throughout we assume a true hypothetical pairwise CG distribution with true interaction parameter matrix $\boldsymbol{S}^\star + \boldsymbol{L}^\star$. Hence, the true model is a fused latent and graphical model whose interactions are decomposed into a group-sparse matrix $\boldsymbol{S}^\star$ of direct interactions among the observed variables and a low-rank matrix $\boldsymbol{L}^\star$ of indirect interactions. Given n samples drawn from this hypothetical true distribution, we try to recover $(\boldsymbol{S}^\star, \boldsymbol{L}^\star)$ by solving Problem (3.7) with suitably chosen regularization parameters. In the following, we denote the solution of Problem (3.7) by $(\boldsymbol{S}_n, \boldsymbol{L}_n)$.

If the estimator given by solving Problem (3.7) succeeds in recovering the true components $\boldsymbol{S}^\star$ and $\boldsymbol{L}^\star$ in the high-dimensional setting, that is, asymptotically and with high probability, then it is called *consistent.* More specifically, we are interested in two types of consistency. The first type is *parametric consistency.* It holds if the errors $\boldsymbol{S}_n - \boldsymbol{S}^\star$ and $\boldsymbol{L}_n - \boldsymbol{L}^\star$ are small at the same time. As in [Chandrasekaran et al., 2012], we measure the size of the errors in the dual norm of the regularizing norm $\gamma\|\boldsymbol{S}\|_{1,2} + \operatorname{tr}(\boldsymbol{L})$ from the objective function. This dual norm is the *γ-norm* defined by

$$\|(\boldsymbol{M}, \boldsymbol{N})\|_\gamma = \max\left\{\gamma^{-1}\|\boldsymbol{M}\|_{\infty,2}, \|\boldsymbol{N}\|\right\}, \quad (\boldsymbol{M}, \boldsymbol{N}) \in \operatorname{Sym}(w) \times \operatorname{Sym}(w),$$

where $w = m + q$ is the dimension of the interaction parameter matrix, $\|\cdot\|_{\infty,2}$ is the $\ell_{\infty,2}$-norm, and $\|\cdot\|$ is the spectral norm. Note that the same γ as in the objective function of Problem (3.7) is used. The second type of consistency is *algebraic consistency.* It holds if $\boldsymbol{S}_n$ has the same group support as $\boldsymbol{S}^\star$ and if $\boldsymbol{L}_n$ retrieves the true rank of $\boldsymbol{L}^\star$.

Consistent recovery is not always possible. A first challenge is controlling the sampling error, which is given by $\nabla\ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star) = \nabla a(\boldsymbol{S}^\star + \boldsymbol{L}^\star) - \hat{\boldsymbol{\Sigma}} = \mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}}$, see Appendix C.2. Here, the expectation is w.r.t. the true model. If the sampling error is small, then it follows from a Taylor expansion that

$$\begin{aligned}\ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star + \boldsymbol{\Delta}) &= \ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star) + \nabla\ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star)^\mathsf{T}\boldsymbol{\Delta} + \frac{1}{2}\,\boldsymbol{\Delta}^\mathsf{T}\nabla^2\ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star)\boldsymbol{\Delta} + R(\boldsymbol{\Delta})\\ &\approx \ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star) + \frac{1}{2}\,\boldsymbol{\Delta}^\mathsf{T}\nabla^2\ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star)\boldsymbol{\Delta} + R(\boldsymbol{\Delta}),\\ &\approx \ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star) + \frac{1}{2}\,\boldsymbol{\Delta}^\mathsf{T}\nabla^2\ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star)\boldsymbol{\Delta}, \end{aligned} \tag{3.8}$$

where the last approximation holds locally since the remainder $R(\boldsymbol{\Delta})$ is small if $\boldsymbol{\Delta}$ is small. The quadratic form in the last line is obviously minimized at $\boldsymbol{\Delta} = \boldsymbol{0}$, which would entail consistent recovery of the compound matrix $\boldsymbol{S}^\star + \boldsymbol{L}^\star$ in a parametric sense. Thus, the likelihood term in Problem (3.7) ensures that the compound matrix $\boldsymbol{S}_n + \boldsymbol{L}_n$ is close to the true compound matrix $\boldsymbol{S}^\star + \boldsymbol{L}^\star$. However, reliable recovery of the compound matrix is only possible if the trade-off parameter λ is not too large. Indeed, the regularization terms should only encourage small adjustments to the algebraic structure of the components. Hence, later we assume an upper bound on the regularization parameter λ.

A second challenge is telling the group-sparse and low-rank components apart. This can be addressed by restricting the analysis to identifiable models, similarly as we did in the analysis of RPCA models in Section 2.2.1 of Chapter 2. Overall, we make assumptions as in [Chandrasekaran et al., 2012] and in [Nussbaum and Giesen, 2019a, 2020a,b]. These assumptions differ depending on the types of observed variables.

3.2.1 An intuitive version of the problem and assumptions

Problem (3.7) can be intuitively understood as minimizing the negative log-likelihood subject to $\boldsymbol{S}$ having a certain group sparsity and $\boldsymbol{L}$ having a certain low rank. As in Section 2.2.1, we use group-sparse and low-rank matrix varieties to formalize these constraints. The only difference to the definitions in Section 2.2.1 lies in the different group structure and that here, we have symmetric matrices. First, the variety of (group-)sparse symmetric matrices with at most s non-zero entries is given as

$$\mathcal{S}(s) = \{\boldsymbol{S} \in \mathrm{Sym}(w) : |\operatorname{gsupp}(\boldsymbol{S})| \leq s\},$$

where

$$\operatorname{gsupp}(\boldsymbol{S}) = \{(i,j) \in [d+q] \times [d+q] : \ \boldsymbol{S}_{ij} \not\equiv \boldsymbol{0}\}$$

is the group support of $\boldsymbol{S}$ (here, $\boldsymbol{S}_{ij}$ is the group of interaction parameters for the i-th and j-th variable of the model, see also Appendix C.2). Second, the variety of matrices with rank at most r is given as

$$\mathcal{L}(r) = \{\boldsymbol{L} \in \mathrm{Sym}(w) : \ \operatorname{rank}(\boldsymbol{L}) \leq r\}.$$

Let us now consider the problem

$$\min_{\boldsymbol{S},\,\boldsymbol{L}\,\in\,\mathrm{Sym}(w)} \ell(\boldsymbol{S}+\boldsymbol{L}) \quad \text{s.t.} \quad \boldsymbol{S} \in \mathcal{S}(|\operatorname{gsupp}(\boldsymbol{S}^\star)|) \ \text{and} \ \boldsymbol{L} \in \mathcal{L}(\operatorname{rank}(\boldsymbol{L}^\star)). \tag{3.9}$$

Similar as the non-convex Problem (2.3) in Chapter 2, this problem is hypothetical since the true sparse and low-rank varieties are not known beforehand. Nevertheless, it offers valuable insights. The important observation is that $(\boldsymbol{S}, \boldsymbol{L})$ can only be (locally) optimal for this intuitive but hypothetical non-convex problem if it satisfies the following optimality condition: The gradient of the negative log-likelihood at $\boldsymbol{S} + \boldsymbol{L}$ (which is the same with respect to $\boldsymbol{S}$ and $\boldsymbol{L}$) is normal to both the group-sparse matrix variety at $\boldsymbol{S}$ and the low-rank matrix variety at $\boldsymbol{L}$. The optimality conditions are visualized in Figure 3.1.

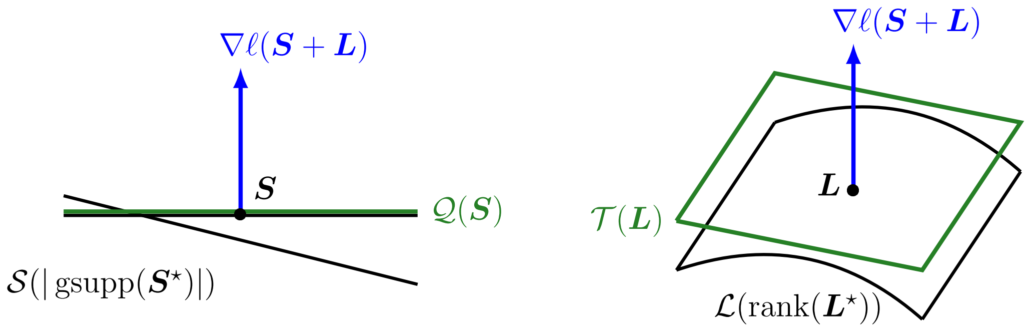
$\nabla\ell(\boldsymbol{S}+\boldsymbol{L})$
$\boldsymbol{S}$
$\mathcal{Q}(\boldsymbol{S})$
$\mathcal{S}(|\operatorname{gsupp}(\boldsymbol{S}^\star)|)$
$\nabla\ell(\boldsymbol{S}+\boldsymbol{L})$
$\boldsymbol{L}$
$\mathcal{T}(\boldsymbol{L})$
$\mathcal{L}(\operatorname{rank}(\boldsymbol{L}^\star))$

Figure 3.1: Illustration of the optimality conditions for Problem (3.9). The gradient of the negative log-likelihood must be normal to the group-sparse and low-rank matrix varieties. Equivalently, the gradient must be normal to the respective tangent spaces $\mathcal{Q}(\boldsymbol{S})$ and $\mathcal{T}(\boldsymbol{L})$.

Stability. For successful recovery, we want the solution of the intuitive Problem (3.9) to be (locally) unique. To guarantee local uniqueness, the optimality condition should be violated at any *slightly perturbed* solution, that is, the gradient at such a perturbed solution should be non-normal to at least one of the varieties. We refer to this property as *stability* (of the solution).

Of particular interest are perturbations in directions of tangent spaces to the respective varieties because the normal spaces to the varieties at tangentially perturbed solutions barely change, if at all. We have already seen the tangent spaces in Section 2.2.1. Here, the tangent spaces take a slightly simpler form since in this chapter we embed the varieties in the set of symmetric matrices $\mathrm{Sym}(w)$. First, the tangent space at a matrix $\boldsymbol{S}$ to the group-sparse matrix variety $\mathcal{S}(|\operatorname{gsupp}(\boldsymbol{S})|)$ is given by

$$\mathcal{Q}(\boldsymbol{S}) = \{\boldsymbol{M} \in \mathrm{Sym}(w) : \ \operatorname{gsupp}(\boldsymbol{M}) \subseteq \operatorname{gsupp}(\boldsymbol{S})\}.$$

Second, a rank-r matrix $\boldsymbol{L}$ is a smooth point of the low-rank matrix variety $\mathcal{L}(r)$ with tangent space

$$\mathcal{T}(\boldsymbol{L}) = \left\{\boldsymbol{U}\boldsymbol{X}^\mathsf{T} + \boldsymbol{X}\boldsymbol{U}^\mathsf{T} : \ \boldsymbol{X} \in \mathbb{R}^{w\times r}\right\},$$

featuring a restricted eigenvalue decomposition $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{E}\boldsymbol{U}^\mathsf{T}$ of $\boldsymbol{L}$, that is, $\boldsymbol{U} \in \mathbb{R}^{w\times r}$ consists of eigenvectors in its columns and $\boldsymbol{E} \in \mathbb{R}^{r\times r}$ is a diagonal matrix of corresponding eigenvalues.

Now certainly, the optimality condition at a tangentially perturbed solution is violated if the gradient at the perturbed solution is tilted in the sense that it is not normal to the varieties anymore. This is the case if the gradient at the perturbed solution has tangential components that are large compared to its components in the respective normal spaces. In the following, we only consider perturbations from the true solution $(\boldsymbol{S}^\star, \boldsymbol{L}^\star)$. The intuition is that stability for this true solution carries over to the solutions of Problems (3.7) and (3.9), provided that they are close to the true solution. For perturbations of the true solution, differentiating Equation (3.8) yields

$$\nabla\ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star + \boldsymbol{\Delta}) - \nabla\ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star) \approx \nabla^2\ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star)\boldsymbol{\Delta} = H^\star\boldsymbol{\Delta}$$

for small $\boldsymbol{\Delta}$. Hence, the Hessian $H^\star = \nabla^2\ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star)$ locally governs the change of the gradient. Therefore, in the following we present conditions on the Hessian that imply that the gradient for (tangentially) perturbed solutions is tilted, see Figure 3.2. A first requirement is that the *minimum gains* of the Hessian $H^\star$ in the respective tangential directions

$$\alpha_{\mathcal{Q}} = \min_{\boldsymbol{\Delta}\in\mathcal{Q},\, \|\boldsymbol{\Delta}\|_{\infty,2}=1} \|P_{\mathcal{Q}}H^\star\boldsymbol{\Delta}\|_{\infty,2}, \quad \alpha_{\mathcal{T},\varepsilon} = \min_{\rho(\mathcal{T},\mathcal{T}')\le\varepsilon}\ \min_{\boldsymbol{\Delta}\in\mathcal{T}',\, \|\boldsymbol{\Delta}\|=1} \|P_{\mathcal{T}'}H^\star\boldsymbol{\Delta}\|$$

should be large since they imply a large tangential component of the gradient at the perturbed solution. Here, we denote projections onto matrix subspaces by P with

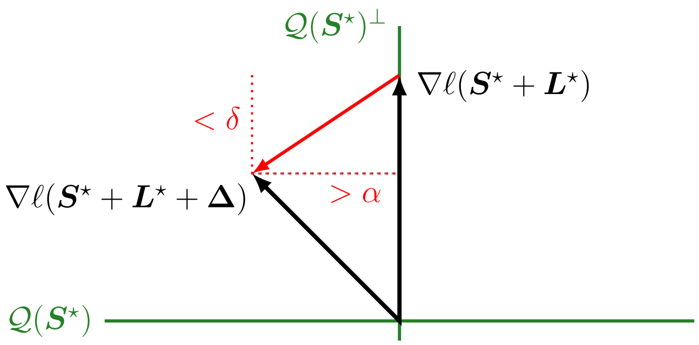
$\mathcal{Q}(\boldsymbol{S}^\star)^\perp$
$\nabla \ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star)$
$< \delta$
$\nabla \ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star + \boldsymbol{\Delta})$
$> \alpha$
$\mathcal{Q}(\boldsymbol{S}^\star)$

Figure 3.2: Tilting of the gradient for a perturbed solution, here in the tangential direction $\boldsymbol{\Delta} \in \mathcal{Q}(\boldsymbol{S}^\star)$. To avoid normality of the gradient at the perturbed solution, the tangential component of the change $\nabla \ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star + \boldsymbol{\Delta}) - \nabla \ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star) \approx \nabla^2 \ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star)\boldsymbol{\Delta}$ of the gradient should be large, and the normal component should be small. The same should hold for perturbations in directions of low-rank tangent spaces that are close to $\mathcal{T}(\boldsymbol{L}^\star)$.

corresponding subscript. Moreover, we set $\mathcal{Q} = \mathcal{Q}(\boldsymbol{S}^\star)$ and $\mathcal{T} = \mathcal{T}(\boldsymbol{L}^\star)$, and given some $\varepsilon > 0$, we considered tangent spaces $\mathcal{T}' \subseteq \operatorname{Sym}(w)$ to the low-rank matrix variety that are close to $\mathcal{T}$ in terms of the twisting

$$\rho(\mathcal{T}, \mathcal{T}') = \max_{\|\boldsymbol{M}\|=1} \|[P_{\mathcal{T}} - P_{\mathcal{T}'}]\,(\boldsymbol{M})\|$$

between these subspaces. This is necessary because the low-rank matrix variety is locally curved such that the tangent spaces for nearby points are often different. Next, we also need the *maximum effects* of the Hessian $H^\star$ in the respective normal directions

$$\delta_{\mathcal{Q}} = \max_{\boldsymbol{\Delta}\in\mathcal{Q},\, \|\boldsymbol{\Delta}\|_{\infty,2}=1} \|P_{\mathcal{Q}^\perp} H^\star \boldsymbol{\Delta}\|_{\infty,2}, \quad \delta_{\mathcal{T},\varepsilon} = \max_{\rho(\mathcal{T},\mathcal{T}')\le\varepsilon} \ \max_{\boldsymbol{\Delta}\in\mathcal{T}',\, \|\boldsymbol{\Delta}\|=1} \|P_{\mathcal{T}'^\perp} H^\star \boldsymbol{\Delta}\|$$

to be small. This is because otherwise the gradient of the negative log-likelihood at the perturbed (true) solution could still be almost normal to the respective varieties. Here, we used the superscript $\perp$ to denote the respective orthogonal/normal spaces.

Note that in our definitions of the minimum gains and maximum effects we used the $\ell_{\infty,2}$- and the spectral norm. These norms are respectively dual to the $\ell_{1,2}$- and the nuclear norm used in the regularization. Eventually, we want to express the stability assumption in the γ-norm which is the dual norm to the joint regularization term in Problem (3.7). This makes a comparison of the $\ell_{\infty,2}$- and the spectral norms on the tangent spaces necessary. The following norm compatibility constants, which are analogous to the ones in Section 2.2.1, serve this purpose:

$$\mu(\mathcal{Q}(\boldsymbol{S})) = \max_{\boldsymbol{M}\in\mathcal{Q}(\boldsymbol{S}):\, \|\boldsymbol{M}\|_{\infty,2}=1} \|\boldsymbol{M}\| \quad \text{and} \quad \xi(\mathcal{T}(\boldsymbol{L})) = \max_{\boldsymbol{N}\in\mathcal{T}(\boldsymbol{L}):\, \|\boldsymbol{N}\|=1} \|\boldsymbol{N}\|_{\infty,2}.$$

Here, $\mathcal{Q}(\boldsymbol{S})$ and $\mathcal{T}(\boldsymbol{L})$ are the tangent spaces at points $\boldsymbol{S}$ and $\boldsymbol{L}$ from the group-sparse matrix variety $\mathcal{S}(|\operatorname{gsupp}(\boldsymbol{S})|)$ and the low-rank matrix variety $\mathcal{L}(\operatorname{rank}(\boldsymbol{L}))$, respectively.

Let us now specify our first main assumption that assures that the ratio of the maximum orthogonal effects and minimum tangential gains is not too large. This makes sure that the gradient at a solution of which a *single* component has been tangentially perturbed is no longer normal to the varieties.

Assumption 1 (Stability)**.** Let $\eta = \max_{i\in[d]} m_i$. Setting $\varepsilon = \xi(\mathcal{T})/(2\eta)$, we assume that

$$\alpha = \min\left\{\alpha_{\mathcal{Q}}, \alpha_{\mathcal{T},\xi(\mathcal{T})/(2\eta)}\right\} > 0 \quad \text{and that} \quad \frac{\delta}{\alpha} < 1,$$

where $\delta = \max\left\{\delta_{\mathcal{Q}}, \delta_{\mathcal{T},\xi(\mathcal{T})/(2\eta)}\right\}$.

To quantify the degree to which Assumption 1 holds, we define $\nu = (1 - \delta/\alpha)/2$. It holds $\nu \in (0, 1/2]$ and the closer ν is to $1/2$, the better. Note that the stability assumption is a generalization of the irrepresentability assumption as, for instance, used by Ravikumar et al. [2011] for Gaussian sparse graphical model selection.

Feasible values for γ. Our next goal is to find values of γ for which we can ensure that the gradient at a solution of which both components have been *simultaneously* perturbed in tangential directions is also not normal to the varieties any longer. The choice of γ is relevant since for simultaneous tangential perturbations we want to measure the maximum orthogonal effects and the minimum tangential gains of the Hessian $H^\star$ in the γ-norm.

Working with the γ-norm requires us to compare the $\ell_{\infty,2}$- and spectral norms. Hence, we can get further insights into the realm of problems for which consistent recovery is possible by taking a look at the norm compatibility constants $\mu(\mathcal{Q}(\boldsymbol{S}^\star))$ and $\xi(\mathcal{T}(\boldsymbol{L}^\star))$. Here, a similar result as Lemma 2.3 holds, where the maximum (group) degree and incoherence are defined in the same way as in Section 2.2.1.

Lemma 3.1. *Let $\boldsymbol{L} \in \mathcal{L}(\operatorname{rank}(L))$ and $\boldsymbol{S} \in \mathcal{S}(|\operatorname{gsupp}(\boldsymbol{S})|)$. Then, the following bounds on the norm compatibility constants hold*

$$\operatorname{coh}(\boldsymbol{L}) \geq 1/(2\eta)\, \xi(\mathcal{T}(\boldsymbol{L})) \quad \textit{and} \quad \operatorname{gdeg}_{\max}(\boldsymbol{S}) \geq \eta^{-1/2}\mu(\mathcal{Q}(\boldsymbol{S})).$$

Proof. The proof follows the lines of the proof of Lemma 2.3. The qualitative differences of the bounds are due to the different group structures, specifically, to prove Lemma 3.1, the norm bound $\|\cdot\|_{\infty,2} \leq \eta\|\cdot\|_\infty$ is used. ■

As in Section 2.2.1, to avoid confusion of $\boldsymbol{S}^\star$ with a low-rank matrix, $\mu(\mathcal{Q}) = \mu(\mathcal{Q}(\boldsymbol{S}^\star)$ should be small since otherwise $\operatorname{gdeg}_{\max}(\boldsymbol{S}^\star)$ is large because of the lower bound from Lemma 3.1 in terms of $\mu(\mathcal{Q})$. Similarly, to avoid confusion of $\boldsymbol{L}^\star$ with a group-sparse matrix, $\xi(\mathcal{T}) = \xi(\mathcal{T}(\boldsymbol{L}^\star))$ should be small since otherwise $\operatorname{coh}(\boldsymbol{L}^\star)$ is large given the lower bound from Lemma 3.1. The fact that both norm compatibility constants should be small is reflected in our second main assumption.

Assumption 2 (γ-feasibility)**.** We assume that

$$\mu(\mathcal{Q})\xi(\mathcal{T}) \le \frac{1}{6}\left(\frac{\alpha\nu}{\beta(2-\nu)}\right)^2,$$

where $\beta = \max\{\beta_{\mathcal{Q}}, \beta_{\mathcal{T}}\}$ with

$$\beta_{\mathcal{Q}} = \max_{\boldsymbol{M}\in\mathcal{Q},\, \|\boldsymbol{M}\|=1} \|H^\star \boldsymbol{M}\| \quad \text{and} \quad \beta_{\mathcal{T}} = \max_{\rho(\mathcal{T},\mathcal{T}') \le \xi(\mathcal{T})/(2\eta)} \; \max_{\boldsymbol{M}\in\mathcal{T}',\, \|\boldsymbol{M}\|_{\infty,2}=1} \|H^\star \boldsymbol{M}\|_{\infty,2}.$$

We introduced the problem-specific constant β since it facilitates the necessary coupling of the $\ell_{\infty,2}$- and the spectral norm when we measure the change of the gradient in the γ-norm for simultaneous perturbations of $(\boldsymbol{S}^\star, \boldsymbol{L}^\star)$ in tangential directions. This is because $\beta_{\mathcal{Q}}$ measures elements from $\mathcal{Q}$ in the spectral norm that is typical for elements from low-rank tangent spaces, and vice versa, $\beta_{\mathcal{T}}$ measures elements from low-rank tangent spaces in the $\ell_{\infty,2}$-norm that is typical for elements from $\mathcal{Q}$. Now, the γ-feasibility assumption implies that the range

$$[\gamma_{\min}, \gamma_{\max}] = \left[\frac{3\beta(2-\nu)\xi(\mathcal{T})}{\nu\alpha}, \frac{\nu\alpha}{2\beta(2-\nu)\mu(\mathcal{Q})}\right]$$

is non-empty. The following proposition shows that for $\gamma \in [\gamma_{\min}, \gamma_{\max}]$ we can bound the minimum gains and the maximum effects of the Hessian $H^\star = \nabla^2 \ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star)$ on the direct sum of the tangent space $\mathcal{Q} = \mathcal{Q}(\boldsymbol{S}^\star)$ and any tangent space $\mathcal{T}'$ close to the true tangent space $\mathcal{T} = \mathcal{T}(\boldsymbol{L}^\star)$. Similarly as the stability assumption does for individual perturbations, these bounds ensure that the gradient of the negative log-likelihood at a simultaneously tangentially perturbed $(\boldsymbol{S}^\star + \boldsymbol{M}, \boldsymbol{L}^\star + \boldsymbol{N})$ with small $\boldsymbol{M} \in \mathcal{Q}$ and small $\boldsymbol{N} \in \mathcal{T}'$ cannot be normal to the varieties any longer.

Proposition 3.2 (Coupled stability)**.** *Suppose that Assumption 1 (stability) and Assumption 2 (γ-feasibility) hold and let $\gamma \in [\gamma_{\min}, \gamma_{\max}]$. Let $\mathcal{T}'$ be a tangent space to the low-rank matrix variety with bounded twisting $\rho(\mathcal{T}, \mathcal{T}') \le \xi(\mathcal{T})/(2\eta)$. Let $\mathcal{J} = \mathcal{Q} \times \mathcal{T}'$. Then,*

(i) *the minimum gain on $\mathcal{J}$ of $H^\star$ restricted to the direct sum $\mathcal{Q} \oplus \mathcal{T}'$ is bounded from below, that is, for all $(\boldsymbol{M}, \boldsymbol{N}) \in \mathcal{J}$ it holds that*

$$\|P_{\mathcal{J}} D H^\star(\boldsymbol{M} + \boldsymbol{N})\|_\gamma \ge \frac{\alpha}{2} \|(\boldsymbol{M}, \boldsymbol{N})\|_\gamma,$$

where $D\colon \mathrm{Sym}(w) \to \mathrm{Sym}(w) \times \mathrm{Sym}(w), \boldsymbol{A} \mapsto (\boldsymbol{A}, \boldsymbol{A})$ is the duplication operator, and

(ii) *the maximum effect on $\mathcal{J}^\perp$ of $H^\star$ restricted to $\mathcal{Q} \oplus \mathcal{T}'$ is bounded from above, that is, for all $(\boldsymbol{M}, \boldsymbol{N}) \in \mathcal{J}$ it holds that*

$$\|P_{\mathcal{J}^\perp} D H^\star(\boldsymbol{M} + \boldsymbol{N})\|_\gamma \le (1-\nu) \|P_{\mathcal{J}} D H^\star(\boldsymbol{M} + \boldsymbol{N})\|_\gamma.$$

The proof of Proposition 3.2 can be found in Appendix C.3.3. We remark that coupled stability implies transversality. To see this, suppose there exists $\mathbf{0} \neq \boldsymbol{A} \in \mathcal{Q} \cap \mathcal{T}'$. Then, choosing $\boldsymbol{M} = \boldsymbol{A}$ and $\boldsymbol{N} = -\boldsymbol{A}$ in Proposition 3.2(a) contradicts the stability assumption because

$$0 = \|P_{\mathcal{J}} DH^\star(\boldsymbol{A} + (-\boldsymbol{A}))\|_\gamma \geq \frac{\alpha}{2} \|(\boldsymbol{A}, -\boldsymbol{A})\|_\gamma = \frac{\alpha}{2} \max\left\{ \frac{\|\boldsymbol{A}\|_{\infty,2}}{\gamma}, \|\boldsymbol{A}\| \right\} > 0$$

since $\alpha > 0$. Hence, we have transversality, that is, $\mathcal{Q} \cap \mathcal{T}' = \{\mathbf{0}\}$.

Gap assumption. We make a final assumption that is necessary for obtaining algebraic consistency. It concerns the smallest-magnitude of the non-zero groups of $\boldsymbol{S}^\star$ given by $s_{\min} = \min_{(i,j) \in \operatorname{gsupp}(\boldsymbol{S}^\star)} \|\boldsymbol{S}^\star_{ij}\|_2$, and it concerns the smallest non-zero eigenvalue $\sigma_{\min}$ of $\boldsymbol{L}^\star$. If they are too small, it will be difficult to recover the true support of $\boldsymbol{S}^\star$ and the true rank of $\boldsymbol{L}^\star$. Hence, we assume a lower bound on both.

Assumption 3 (Gap)**.** We require that

$$s_{\min} \geq \frac{C_S \lambda}{\mu(\mathcal{Q})} \quad \text{and} \quad \sigma_{\min} \geq C_L \max\left\{ \frac{\eta}{\xi(\mathcal{T})^2}, 1 \right\} \lambda,$$

where C_S and C_L are problem-specific constants that are defined in Appendix C.3.2.

3.2.2 Consistency of pairwise sparse + low-rank models

In this section, we state a number of consistency results for pairwise sparse + low-rank models. Recall that $w = m + q$ is the number of observed variables (counting the number of interacting indicator variables for discrete variables). As before, we denote the tangent space to the variety of symmetric sparse matrices at $\boldsymbol{S}^\star$ by $\mathcal{Q} = \mathcal{Q}(\boldsymbol{S}^\star)$ and the one to the variety of symmetric low-rank matrices at $\boldsymbol{L}^\star$ by $\mathcal{T} = \mathcal{T}(\boldsymbol{L}^\star)$. In this section, we use some problem-specific constants $C_1, \dots, C_5$ whose exact definitions can be found in Appendix C.3.2.

For consistent recovery, the sample must represent the underlying distribution well. This is the case if the sampling error $\boldsymbol{\Sigma}^\star - \hat{\boldsymbol{\Sigma}}$ of the second-moment matrix is small, where $\boldsymbol{\Sigma}^\star = \mathbb{E}[\boldsymbol{\Sigma}]$ is the expected and $\hat{\boldsymbol{\Sigma}}$ is the empirical second-moment matrix. Remember that the gradient of the negative log-likelihood coincides with the sampling error, that is, it holds $\nabla \ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star) = \boldsymbol{\Sigma}^\star - \hat{\boldsymbol{\Sigma}}$. Hence, first we explicitly assume that the sample is 'good' in the sense that its sampling error is small. For such samples, the following theorem shows that the solution to Problem (3.7) is consistent.

Theorem 3.3. *Let $(\boldsymbol{S}^\star, \boldsymbol{L}^\star)$ such that the stability, γ-feasibility, and the gap assumption (for λ as chosen below) are satisfied. Suppose that we observed samples*

$$(\boldsymbol{x}^{(1)}, \boldsymbol{y}^{(1)}), \ldots, (\boldsymbol{x}^{(n)}, \boldsymbol{y}^{(n)}) \in \mathcal{X} \times \mathcal{Y} = \prod_{i=1}^{d} \{0, \ldots, m_i\} \times \mathbb{R}^q$$

drawn from a pairwise CG model with interaction parameter matrix $\boldsymbol{S}^\star + \boldsymbol{L}^\star$. Moreover, let $\lambda \leq \min\{C_1, C_2\,\xi(\mathcal{T})\}$ and $\gamma \in [\gamma_{\min}, \gamma_{\max}]$. Then, if the gradient at the true parameters $\boldsymbol{S}^\star + \boldsymbol{L}^\star$ is bounded in the γ-norm via

$$\|D\nabla\ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star)\|_\gamma = \left\|D(\boldsymbol{\Sigma}^\star - \hat{\boldsymbol{\Sigma}})\right\|_\gamma \leq C_3\lambda,$$

the solution $(\boldsymbol{S}_n, \boldsymbol{L}_n)$ to the convex Problem (3.7) *with regularization parameters λ and γ exists and is unique. Furthermore, it is*

a) parametrically consistent by virtue of satisfying $\|(\boldsymbol{S}_n - \boldsymbol{S}^\star, \boldsymbol{L}_n - \boldsymbol{L}^\star)\|_\gamma \leq C_4\lambda$ and

b) algebraically consistent, that is, the group supports of $\boldsymbol{S}_n$ and $\boldsymbol{S}^\star$ are the same, and the ranks of $\boldsymbol{L}_n$ and $\boldsymbol{L}^\star$ coincide.

We make a few comments to explain this theorem: First, we assumed an upper bound on λ to make sure that the shrinkage effects on the solution caused by the regularization terms from the objective function are not too large. Second, we did not assume a lower bound for λ, but for small λ it will be difficult to achieve algebraic consistency. This is reflected in the assumed bound on the sampling error that we formulated in terms of λ. Here, if λ is small, then the number of (random) samples that are required so that the bound actually holds with high probability is large. This number of samples depends on the types of observed variables (binary-only, quantitative-only, both binary and quantitative).

The proof of Theorem 3.3 in Appendix C.3.2 - C.3.6 generalizes the proof in [Chandrasekaran et al., 2012] for models with observed Gaussian, the proof in [Nussbaum and Giesen, 2019b] for models with observed binary, and the proof in [Nussbaum and Giesen, 2020b] for models with observed binary and quantitative variables. The proof is a version of the primal-dual witness technique, which has originally been used for the Lasso [Wainwright, 2009] and later for sparse graphical model selection, see, for example, [Ravikumar et al., 2010]. It proceeds by first restricting Problem (3.7) to a (non-convex) correct model set $\mathcal{M}$ chosen in a way such that any solution $(\boldsymbol{S}_\mathcal{M}, \boldsymbol{L}_\mathcal{M})$ to the restricted problem is algebraically and parametrically consistent. The non-convexity is due to a rank constraint, which is subsequently linearized by replacing it with a tangent-space constraint to the low-rank matrix variety at a fixed solution $(\boldsymbol{S}_\mathcal{M}, \boldsymbol{L}_\mathcal{M})$. Then, it is shown that the solution to the linearized problem is unique and coincides with $(\boldsymbol{S}_\mathcal{M}, \boldsymbol{L}_\mathcal{M})$. Finally, it is shown that the original Problem (3.7) is also solved by the same consistent solution $(\boldsymbol{S}_\mathcal{M}, \boldsymbol{L}_\mathcal{M})$. Moreover, this solution is shown to be *strictly dual feasible*, which ensures that there cannot be other solutions. A more detailed outline of the proof can be found in Appendix C.3.1.

In the following, we derive explicit regimes for the different types of variables. They will allow us to see that Theorem 3.3 can be used to determine the asymptotic behavior when the number of variables q and/or d, the number of samples n, and the number of latent variables r grow. For that, we make a specific choice for λ given by

$$\lambda = \lambda_{n,f} = \frac{C_5}{\xi(\mathcal{T})}\sqrt{\frac{f}{n}},$$

where f may functionally depend on q if quantitative observed variables are present, and f may also depend on d and m in the presence of discrete observed variables. We define explicit expressions of f for the different types of observed variables below, where f can only grow in q, d, and m. Here, provided that the lower bound

$$n \geq \frac{C_5^2\, f}{\xi(\mathcal{T})^2 \min\{C_1, C_2\, \xi(\mathcal{T})\}^2} \tag{3.10}$$

holds, then the assumption $\lambda_{n,f} \leq \min\{C_1, C_2\, \xi(\mathcal{T})\}$ of Theorem 3.3 is satisfied. We will respectively choose the scaling f such that also the bound on the gradient/sampling error from Theorem 3.3 holds with high probability, assuming that the number of samples n satisfies a lower bound with the same dependence on (q, d, m) as in (3.10). Note that the dependence on (q, d, m) in (3.10) is completely specified by the choice of f. Hence, the lower bound (3.10) and the scaling f determine the sample complexity for the respective scenarios: Larger f implies a stronger lower bound on n. Let us further discuss the asymptotic behavior obtained from Theorem 3.3 with the choice $\lambda = \lambda_{n,f}$. First, observe that for $n \to \infty$ and fixed (q, d, m), that is, fixed f, it follows that

$$\|(\boldsymbol{S}_n - \boldsymbol{S}^\star, \boldsymbol{L}_n - \boldsymbol{L}^\star)\|_\gamma \leq C_4\, \lambda_{n,f} \to 0.$$

This means that asymptotically the errors become zero. Second, for a larger number of variables, f is larger. Hence, achieving the same error bound requires more samples. This is natural because the number of parameters to be estimated is also larger.

Now, the following corollaries state the scalings f for the special cases with quantitative-only, discrete-only (binary-only), and both discrete and quantitative observed variables. All corollaries use $\lambda = \lambda_{n,f}$ as a regularization parameter (for the respectively chosen scaling f) and assume a lower bound on n as in (3.10). The first result matches the one in [Chandrasekaran et al., 2012].

Corollary 3.4 (Consistency Gaussian model). *Let there only be observed quantitative, Gaussian variables ($w = q$), that is, points $\boldsymbol{y}^{(1)}, \dots, \boldsymbol{y}^{(n)} \in \mathbb{R}^q$ are observed. Let $f = f(q) = Cq$ for some constant C (defined in the proof). Then, under the assumptions of Theorem 3.3 with $\lambda = \lambda_{n,f(q)}$, the claims of Theorem 3.3 (parametric and algebraic consistency) hold with probability at least* $1 - 2\exp(-q)$.

Corollary 3.5 (Consistency discrete model)**.** *Let there only be observed discrete variables ($w = m$) such that the sample consists of points $\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(n)} \in \mathcal{X} = \prod_{i=1}^{d}\{0, \dots, m_i\}$. Moreover, let $\kappa \geq 1$. Let the scaling be given by $f = f(d, m) = C\kappa d \log m$ for some constant C (defined in the proof). Then, under the assumptions of Theorem 3.3 with $\lambda = \lambda_{n,f(d,m)}$, the claims of Theorem 3.3 hold with probability at least $1 - m^{-\kappa}$.*

Note that this second result encompasses the result from [Nussbaum and Giesen, 2019a] for the Ising model as a special case since for binary variables Corollary 3.5 simplifies given that $d = m$. The third and final result concerns models with mixed discrete and continuous variables.

Corollary 3.6 (Consistency mixed model)**.** *Let $\kappa > 0$ and let the scaling be given by $f(w) = C(1 + \kappa/2)w^2 \log w$, where $w = q + m$ and C is a constant (defined in the proof). Then, under the assumptions of Theorem 3.3 with $\lambda = \lambda_{n,f(w)}$, the claims of Theorem 3.3 hold with probability at least $1 - 2w^{-\kappa}$.*

The proofs of the corollaries can be found in Appendix C.3.7. They bound the sampling error $\boldsymbol{\Sigma}^\star - \hat{\boldsymbol{\Sigma}}$ in the spectral norm (or rather $D(\boldsymbol{\Sigma}^\star - \hat{\boldsymbol{\Sigma}})$ in the γ-norm as required by Theorem 3.3) for the respective scenarios. Afterwards they apply Theorem 3.3. Here, we just note that the different scalings and sample complexities are due to the fact that bounding the sampling error is more challenging for non-Gaussian distributions since the strong Gaussian tail bounds cannot be used. Indeed, if there are only observed binary variables, the resulting random vector is not even sub-Gaussian, but at least it is bounded. The situation is even more complicated if there are observed discrete *and* quantitative variables. This is because in this case the distribution is neither sub-Gaussian nor bounded.

Note that we can only assume in theory that the constants from the theorem and its corollaries are known—in practice, they cannot be computed, particularly since the true pair $(\boldsymbol{S}^\star, \boldsymbol{L}^\star)$ is unknown. However, a good starting point for choosing regularization parameters is trying $\lambda = k\sqrt{f/n}$ for different values of k. Overall, the two-dimensional search space of the regularization parameters can, for example, be traversed by using grid search, random search, or even more sophisticated methods such as Benson's algorithm [Giesen et al., 2019b]. We present the latter method in Section 3.5.

3.3 Solvers

3.3.1 ADMM algorithm with proximal-gradient steps

Let us first consider a reformulation of Problem (3.7)

$$\min_{\boldsymbol{\Theta}\,:\,\Lambda[\boldsymbol{\Theta}]\succ \boldsymbol{0},\, \boldsymbol{W}} \quad \ell(\boldsymbol{\Theta}) + \varphi(\boldsymbol{W}) \quad \text{s.t.} \quad \boldsymbol{\Theta} = [\boldsymbol{I}, \boldsymbol{I}]\boldsymbol{W} \tag{3.11}$$

where we grouped $\boldsymbol{W} = (\boldsymbol{S}, \boldsymbol{L})$ into one variable, $\boldsymbol{I}$ is the $(w \times w)$-identity matrix, and

$$\varphi(\boldsymbol{W}) = \alpha\|\boldsymbol{S}\|_{1,2} + \beta\,\mathrm{tr}(\boldsymbol{L}) + \chi[\boldsymbol{L} \succeq \boldsymbol{0}].$$

Here, χ is the indicator function that takes the value zero if the condition is satisfied, and infinity otherwise. The regularization parameters of the problem are $\alpha, \beta > 0$.

Let $\kappa > 0$ and let $\boldsymbol{\Phi}$ be the dual variables for the constraint $\boldsymbol{\Theta} = [\boldsymbol{I}, \boldsymbol{I}]\boldsymbol{W}$. First, the algorithm is initialized with variables $(\boldsymbol{\Theta}^0, \boldsymbol{W}^0, \boldsymbol{\Phi}^0)$, for example, $\boldsymbol{\Theta}^0 = -\boldsymbol{I}$, $\boldsymbol{W}^0 = \boldsymbol{\Phi}^0 = \boldsymbol{0}$ is a feasible starting point. Then, the ADMM updates for $k \geq 1$ are given by

$$\begin{cases} \boldsymbol{\Theta}^{k+1} & = \arg\min_{\boldsymbol{\Theta}\,:\,\Lambda[\boldsymbol{\Theta}]\succ\boldsymbol{0}} \ell(\boldsymbol{\Theta}) - \left\langle \boldsymbol{\Phi}^k, \boldsymbol{\Theta} - [\boldsymbol{I}, \boldsymbol{I}]\boldsymbol{W}^k \right\rangle + \frac{1}{2\kappa}\left\|\boldsymbol{\Theta} - [\boldsymbol{I}, \boldsymbol{I}]\boldsymbol{W}^k\right\|_F^2, \\ \boldsymbol{W}^{k+1} & = \arg\min_{\boldsymbol{W}} \varphi(\boldsymbol{W}) - \left\langle \boldsymbol{\Phi}^k, \boldsymbol{\Theta}^{k+1} - [\boldsymbol{I}, \boldsymbol{I}]\boldsymbol{W} \right\rangle + \frac{1}{2\kappa}\left\|\boldsymbol{\Theta}^{k+1} - [\boldsymbol{I}, \boldsymbol{I}]\boldsymbol{W}\right\|_F^2, \\ \boldsymbol{\Phi}^{k+1} & = \boldsymbol{\Phi}^k - \kappa^{-1}(\boldsymbol{\Theta}^{k+1} - [\boldsymbol{I}, \boldsymbol{I}]\boldsymbol{W}^{k+1}), \end{cases}$$

which is equivalent to

$$\begin{cases} \boldsymbol{\Theta}^{k+1} & = \arg\min_{\boldsymbol{\Theta}\,:\,\Lambda[\boldsymbol{\Theta}]\succ\boldsymbol{0}} \ell(\boldsymbol{\Theta}) + \frac{1}{2\kappa}\left\|\boldsymbol{\Theta} - [\boldsymbol{I}, \boldsymbol{I}]\boldsymbol{W}^k - \kappa\boldsymbol{\Phi}^k\right\|_F^2, \\ \boldsymbol{W}^{k+1} & = \arg\min_{\boldsymbol{W}} \varphi(\boldsymbol{W}) + \frac{1}{2\kappa}\left\|\boldsymbol{\Theta}^{k+1} - [\boldsymbol{I}, \boldsymbol{I}]\boldsymbol{W} - \kappa\boldsymbol{\Phi}^k\right\|_F^2, \\ \boldsymbol{\Phi}^{k+1} & = \boldsymbol{\Phi}^k - \kappa^{-1}(\boldsymbol{\Theta}^{k+1} - [\boldsymbol{I}, \boldsymbol{I}]\boldsymbol{W}^{k+1}). \end{cases} \tag{3.12}$$

In the following, we discuss how the optimization problems that appear in the individual updates can be solved (or at least approximately solved).

The first update. The first update is the proximal mapping of the likelihood:

$$\min_{\boldsymbol{\Theta}\,:\,\Lambda[\boldsymbol{\Theta}]\succ\boldsymbol{0}} \ell(\boldsymbol{\Theta}) + \frac{1}{2\kappa}\|\boldsymbol{\Theta} - \boldsymbol{Z}\|_F^2,$$

where $\boldsymbol{Z} = [\boldsymbol{I}, \boldsymbol{I}]\boldsymbol{W}^k + \kappa\boldsymbol{\Phi}^k$. For purely Gaussian models, the zero-mean Gaussian negative log-likelihood is given by

$$\ell(\boldsymbol{\Theta}) = -\log\det\boldsymbol{\Theta} + \langle\boldsymbol{\Theta}, \hat{\boldsymbol{\Sigma}}\rangle + \chi[\boldsymbol{\Theta} \succ \boldsymbol{0}]$$

with empirical covariance matrix $\hat{\boldsymbol{\Sigma}}$. In this case, the proximal operator has the solution

$$\arg\min_{\boldsymbol{\Theta}\,:\,\Lambda[\boldsymbol{\Theta}]\succ\boldsymbol{0}} \ell(\boldsymbol{\Theta}) + \frac{1}{2\kappa}\|\boldsymbol{\Theta} - \boldsymbol{Z}\|_F^2 = \boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\gamma})\boldsymbol{U}^\mathsf{T},$$

where

$$\gamma_i = -\frac{\sigma_i}{2} + \sqrt{\frac{\sigma_i^2}{4} + \kappa}, \qquad \text{for } i = 1, \ldots, q$$

and $\boldsymbol{U}\,\mathrm{diag}(\boldsymbol{\sigma})\boldsymbol{U}^\mathsf{T}$ is a singular value decomposition of $\kappa\hat{\boldsymbol{\Sigma}} - \boldsymbol{Z}$. The solution can be derived from the first-order optimality condition, see [Ma et al., 2013] for the details.

The situation is more complicated in the presence of discrete variables. In this case, an iterative optimization algorithm needs to be applied. Most iterative algorithms, like the Broyden-Fletcher-Goldfarb-Shanno (*BFGS*) algorithm [Fletcher, 2013] and variants thereof use at least first-order information, that is, the gradient of the objective function. Computing the likelihood function can already be computationally expensive because sums over the discrete states in $\mathcal{X}$ must be evaluated. Moreover, using the derivative often causes numerical instabilities (for example, the derivative of $\log\det(\boldsymbol{\Theta})$ is $\boldsymbol{\Theta}^{-1}$, which can easily lead to bad numerical condition). Because of the aforementioned problems, it is often more practical to replace the likelihood by a more computationally tractable pseudo-likelihood. We introduce it in the next section. Nevertheless, as we will see, even the proximal operator of the pseudo-likelihood does in general not have a closed-form solution, making the use of an iterative solver necessary.

The second update. In the problem from the second update in (3.12), the components of $\boldsymbol{W}$ are coupled in the quadratic Frobenius-norm term. With this coupling the proximal operator for $\boldsymbol{W}$ is hard to solve. Instead it has been suggested in [Ma et al., 2013] to solve a step of a proximal-gradient method, that is, for $\tau > 0$ one solves

$$\min_{\boldsymbol{W}} \varphi(\boldsymbol{W}) + \frac{1}{2\kappa\tau} \left\| \boldsymbol{W} - \left(\boldsymbol{W}^k + \tau[\boldsymbol{I}\ \boldsymbol{I}]^\top \left(\boldsymbol{\Theta}^{k+1} - [\boldsymbol{I}, \boldsymbol{I}]\boldsymbol{W}^k - \kappa\boldsymbol{\Phi}^k \right) \right) \right\|_F^2 .$$

Now, in this problem, the components $\boldsymbol{S}$ and $\boldsymbol{L}$ are separable. Consequently the proximal-gradient step reduces to solving two proximal operators, namely the one of the $\ell_{1,2}$-norm and the one of the nuclear norm. We have already seen their solutions in Section 2.3. Hence, the first update is

$$\begin{aligned} \boldsymbol{S}^{k+1} &= \arg\min_{\boldsymbol{S}}\ \alpha\|\boldsymbol{S}\|_{1,2} + \frac{1}{2\kappa\tau} \left\| \boldsymbol{S} - \left(\boldsymbol{S}^k + \tau\boldsymbol{G}^k \right) \right\|_F^2 \\ &= \mathrm{gShrink}(\boldsymbol{S}^k + \tau\boldsymbol{G}^k, \alpha\kappa\tau), \end{aligned}$$

where the group shrinkage operator is as in Equation (2.5) (after an easy adaptation of the group structure). The second update is

$$\begin{aligned} \boldsymbol{L}^{k+1} &= \arg\min_{\boldsymbol{L}}\ \beta\,\mathrm{tr}(\boldsymbol{L}) + \chi[\boldsymbol{L} \succeq \boldsymbol{0}] + \frac{1}{2\kappa\tau} \left\| \boldsymbol{L} - \left(\boldsymbol{L}^k + \tau\boldsymbol{G}^k \right) \right\|_F^2 \boldsymbol{L}^{k+1} \\ &= \boldsymbol{U} \max\{\boldsymbol{E} - \beta\kappa\tau, \boldsymbol{0}\}\boldsymbol{U}^\top, \end{aligned}$$

where $\boldsymbol{G}^k = \boldsymbol{\Theta}^{k+1} - \boldsymbol{S}^k - \boldsymbol{L}^k - \kappa\boldsymbol{\Phi}^k$ and $\boldsymbol{L}^k + \tau\boldsymbol{G}^k = \boldsymbol{U}\boldsymbol{E}\boldsymbol{U}^\top$ is an eigenvalue decomposition with eigenvectors in $\boldsymbol{U}$ and diagonal matrix $\boldsymbol{E}$. Note that the maximum in the last equation is to be understood element-wise.

Convergence. In the original work, Ma et al. [2013] showed convergence of the original proximal-gradient based ADMM algorithm for Gaussian fused latent and graphical models (compare Theorem A.2 in [Ma et al., 2013]). This convergence is independent from the initialization. The proof of convergence has not yet been transferred to our setting, however, it should be possible to generalize the proof of convergence. In practice, we observe convergence of the proposed algorithm.

3.3.2 Pseudo-likelihood

The computation of the standard likelihood function involves a sum over all discrete states for obtaining the normalization constant. This has a high computational cost and often prohibits learning larger models using standard likelihood. Besag [1975] introduced an alternative pseudo-likelihood that is based on conditional probabilities of, respectively, one variable given all the others. Using the pseudo-likelihood has since been a common technique in the estimation of sparse graphical models for discrete variables, see [Jalali et al., 2011; Lee and Hastie, 2015; Ravikumar et al., 2011]. Chen et al. [2018] also used the pseudo-likelihood for estimating binary fused latent and graphical models. It is generally believed that likelihood and pseudo-likelihood behave similarly in applications, see, for example, [Mozeika et al., 2014].

In the following, we present the pseudo-likelihoods for pairwise CG distributions

$$p(\boldsymbol{x}, \boldsymbol{y}) = \exp\left\{\frac{1}{2}(\overline{\boldsymbol{x}}, \boldsymbol{y})^\intercal \boldsymbol{\Theta}(\overline{\boldsymbol{x}}, \boldsymbol{y}) - a(\boldsymbol{\Theta})\right\}, \quad (\boldsymbol{x}, \boldsymbol{y}) \in \prod_{i=1}^{d}\{0, \dots, m_i\} \times \mathbb{R}^q,$$

see also Model (3.6). Remember that the matrix $\boldsymbol{\Theta}$ is group structured, where the groups of discrete-discrete interactions are $(q_{ij;kl})_{k\in[m_i], l\in[m_j]} \in \mathbb{R}^{m_i\times m_j}$ for $i, j \in [d]$, the groups of quantitative-discrete interactions are $(\rho_{si;k})_{k\in[m_i]} \in \mathbb{R}^{m_i}$ for $i \in [d]$ and $s \in [q]$, and the parameters λ_{st} for $s, t \in [q]$ describe pairwise interactions between two quantitative variables, that is, the groups of quantitative-quantitative interactions consist of only single elements. Note that we omitted univariate parameters for the continuous variables in the model definition above (we did so in general in this thesis to ease the theoretical analysis). However, they can be included easily in a practical implementation.

Assume that n data points $(\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)}) \in \mathcal{X} \times \mathcal{Y} = \prod_{i=1}^{d}\{0, \dots, m_i\} \times \mathbb{R}^q$, $k = 1, \dots, n$ have been observed. The negative pseudo log-likelihood is given by

$$\ell_p(\boldsymbol{\Theta}) = -\sum_{k=1}^{n}\left(\sum_{i=1}^{d}\log p(x_i = x_i^{(k)}|\boldsymbol{x}_{-i}^{(k)}, \boldsymbol{y}^{(k)}) + \sum_{s=1}^{q}\log p(y_s = y_s^{(k)}|\boldsymbol{x}^{(k)}, \boldsymbol{y}_{-s}^{(k)})\right), \tag{3.13}$$

where the subscript $-i$ is used to denote the omission of the i-th component of the vector. Note that we use the *negative* pseudo-log-likelihood since this allows us to write down convex minimization problems, and we use the *log* versions since the sum of log terms is computationally more stable than large products.

Let us now take a look at the two types of node conditional distributions that appear in Equation (3.13), compare [Lee and Hastie, 2015].

First, the node conditional distribution of a discrete variable is given by

$$p(x_i = k | \boldsymbol{x}_{-i}, \boldsymbol{y}) = \frac{\exp\left(q_{ii;kk} + \sum_{j:j\neq i} q_{ij;kx_j} + \sum_{s=1}^{q} \rho_{si;k}\, y_s\right)}{\sum_{l\in[m_i]} \exp\left(q_{ii;ll} + \sum_{j:j\neq i} q_{ij;lx_j} + \sum_{s=1}^{q} \rho_{si;l}\, y_s\right)}.$$

Observe the similarity to multinomial logistic models (multi-class classification). Second, the node conditional distribution of a quantitative, conditional Gaussian variable is the univariate Gaussian distribution given by

$$p(y_s|\boldsymbol{x}, \boldsymbol{y}_{-s}) = \frac{\lambda_{ss}}{\sqrt{2\pi}} \exp\left(-\frac{\lambda_{ss}}{2}\left(y_s - \frac{\mu_s(\boldsymbol{x}, \boldsymbol{y}_{-s})}{\lambda_{ss}}\right)^2\right),$$

where

$$\mu_s(\boldsymbol{x}, \boldsymbol{y}_{-s}) = \sum_{i=1}^{d} \rho_{si;x_i} - \sum_{s\neq t} \lambda_{st} y_t = \sum_{i=1}^{d} \rho_{si;x_i} - \frac{1}{2}\sum_{t:t\neq s} (\lambda_{st} + \lambda_{ts})\, y_t$$

is a regression term for the mean.

Discussion of computational efficiency. An evaluation of the pseudo-likelihood avoids the costly computation of normalization constants. On the other hand, computing the standard likelihood only requires the sufficient statistics of the data as input, that is, it is sufficient to know the relevant empirical moments of the data (for example, the sample mean and the sample second-moment matrix). The empirical moments often pose a significant reduction of the data. In contrast, each evaluation of the pseudo-likelihood requires to sum over the data points. This shows that in the case of many observed samples and only a few discrete variables, computing the standard likelihood may still be more computationally efficient. Nevertheless, if any discrete variables are present, we do not use the standard likelihood because of numerical problems that are imminent in the computation of gradients (as we mentioned in the previous section). Indeed, likelihood optimization does not have closed-form solutions when learning pairwise models that involve discrete variables, hence the gradients are needed to make the application of standard (first-order) optimization methods possible.

Unfortunately, also pseudo-likelihood optimization does not have closed-form solutions. Particulary, using the pseudo-likelihood in the context of the proximal-gradient based ADMM solver that we introduced in the previous section comes with the disadvantage that its proximal mapping

$$\min_{\boldsymbol{\Theta}\,:\,\Lambda[\boldsymbol{\Theta}]\succ \boldsymbol{0}} \ \ell_p(\boldsymbol{\Theta}) + \frac{1}{2\kappa}\left\|\boldsymbol{\Theta} - \boldsymbol{Z}\right\|_F^2 \tag{3.14}$$

does not permit closed-form solutions (non-regarding of the types of observed variables). Hence, this proximal mapping must be solved with an iterative solver. Luckily, pseudo-likelihood optimization generally suffers from fewer numerical issues. Nevertheless, invoking an iterative solver for the proximal mapping given by Problem (3.14) makes each iteration of the proximal-gradient based ADMM solver given by the updates in (3.12) rather costly. With some tricks it is possible to boost the performance to some extent, for example, by warm starting the solver for Problem (3.14) with the solution from the previous iteration, or even solving Problem (3.14) only with a moderate accuracy. Hence, using the proximal-gradient based ADMM algorithm with updates as in (3.12) allows for solving problems with several hundreds of variables, or even a few thousands, in a considerable amount of time. Of course, this also depends on the available compute power.

3.3.3 Implementation in Python: the cgmodsel package

The algorithms for learning fused latent and graphical models presented in this section are implemented in the Python package CGMODSEL (the name is derived from CG model selection), see `https://github.com/franknu/cgmodsel`. An older version of this package has also been published alongside our article [Nussbaum and Giesen, 2020b]. The CGMODSEL package contains solvers for estimating the parameters of different conditional Gaussian distributions. Beyond fused latent and graphical models, also solvers for sparse graphical models are implemented (mostly ADMM based). Moreover, the package provides basic maximum-a-posteriori (MAP) estimators.

A documentation is provided alongside the package. Here, we just give a brief overview. There are two main components: *models* and *solvers*. Models are implemented as Python classes. Each model is characterized by a set of parameters. Most models, such as, sparse graphical models and fused latent and graphical models, provide methods for visualizing the parameters. Besides, all kinds of utility are implemented, for example, conversion methods between different models and parametrizations. Solvers are also implemented as Python classes. They handle training data, and they provide an interface to set regularization parameters and to call methods for fitting the model. The correctness of the solvers has been verified by unit tests that compare the solutions to reference solutions from external general purpose solvers, such as, the symmetric cone solver SeDuMi [Sturm, 1999].

3.4 Experiments

We solve Problem (3.7) using the proximal-gradient based ADMM algorithm from the previous section. For computational efficiency, we replace the likelihood by the pseudo-likelihood in our experiments. Mozeika et al. [2014] argued that pseudo-likelihood and likelihood behave similarly. The experiments in this section are based on the work [Nussbaum and Giesen, 2020a].

3.4.1 Synthetic data

Here, to verify experimentally that consistent recovery is possible, we generate synthetic data from discrete fused latent and graphical models using Gibbs sampling, see Casella and George [1992]. For the experiments, we use discrete variables that take three values, that is, $\mathcal{X}_i = \{0, 1, 2\}$ for all variables. We consider four models with $d = 36$ variables, where the direct interactions $\boldsymbol{S}^\star$ adhere to either chain or grid graphical model structures (compare Figure 3.3), and the number of latent variables is either one or two.

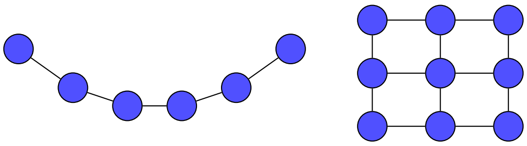

Figure 3.3: Chain (left) and grid (right) graphical model structures. Edges respectively correspond to groups of parameters.

Our goal is to test the influence of the maximum group degree $\operatorname{gdeg}_{\max}(\boldsymbol{S}^\star)$ of $\boldsymbol{S}^\star$ and the incoherence $\operatorname{coh}(\boldsymbol{L}^\star)$ of $\boldsymbol{L}^\star$ on recovery rates. Here, the assumptions of Theorem 3.3 are stated in terms of the more technical constants $\mu(\mathcal{Q}(\boldsymbol{S}^\star))$ and $\xi(\mathcal{T}(\boldsymbol{L}^\star))$, particularly Assumption 2 requires the product of these constants to be small. However, it holds

$$\xi(\mathcal{T}(\boldsymbol{L}^\star))\mu(\mathcal{Q}(\boldsymbol{S}^\star)) \leq \eta^{3/2} \operatorname{gdeg}(\boldsymbol{S}^\star) \operatorname{coh}(\boldsymbol{L}^\star).$$

by Lemma 3.1. It is easy to show that the theoretical results also hold if maximum group degree and incoherence are small (note the similarity to how we obtained Corollary 2.7 in Chapter 2). Observe that the maximum group degree is two for the chain and four for the grid model. Moreover, in our experiments, we set the probability of an interaction between any latent and any observed variable to be non-zero to 95%. This ensures that the low-rank effect of the latent variables is spread-out and thus that $\boldsymbol{L}^\star$ is incoherent. However, $\boldsymbol{L}^\star$ will be less incoherent for a growing number of latent variables.

For each model, we sample all of its parameters randomly. More specifically, we sample the latent-observed interaction parameters uniformly from $[-0.5, -0.2] \cup [0.2, 0.5]$ and the parameters for the non-zero groups of $\boldsymbol{S}^\star$ from $[-1.5, -0.5] \cup [0.5, 1.5]$. Then, for each model we test the asymptotic behavior by generating 20 datasets with $kd \log m$ samples (rounded to the nearest integer) for selected ratios $k \in [1, 50]$. Our choice of regularization parameters is guided by Corollary 3.5 and fixed for all models ($\lambda = 1/50\,\sqrt{d \log m/n}$, $\gamma = 10$). Finally, for each model and ratio k, we record the average percentage of correctly identified non-zero groups, that is, edges in the conditional independence graph. For that, we employ the criteria of *recall* and *precision*, where recall $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$ and precision $= \mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$. Here, TP is the number of correctly identified edges (true positives), FN is the number of undetected edges (false negatives), and FP is the number of edges that were mistakenly detected as edges (false positives). Likewise, we record the absolute rank difference $|\operatorname{rank}(\hat{\boldsymbol{L}}) - \operatorname{rank}(\boldsymbol{L}^\star)|$, averaged over the 20 trials.

The results are shown in Figure 3.4. Recovery of edges and rank requires relatively few samples for the one-latent-variable chain and grid models. Slightly more samples are required to recover the rank of the grid model. This is due to the larger maximum group degree of the grid models compared to the chain models. Next, for the two-latent-variable chain model considerably more samples are necessary for successful recovery—because the underlying low-rank matrix is less incoherent. Our observations back the intuition that for more incoherent $\boldsymbol{L}^\star$ and smaller maximum group degree of $\boldsymbol{S}^\star$, the group-sparse and low-rank components can be confused less easily. This is supported even more by the recovery results for the two-latent-variable grid model, where the fact that both the maximum group degree and the coherence are larger is reflected in significantly worse recovery results. Nevertheless, overall the results show that consistent recovery is possible.

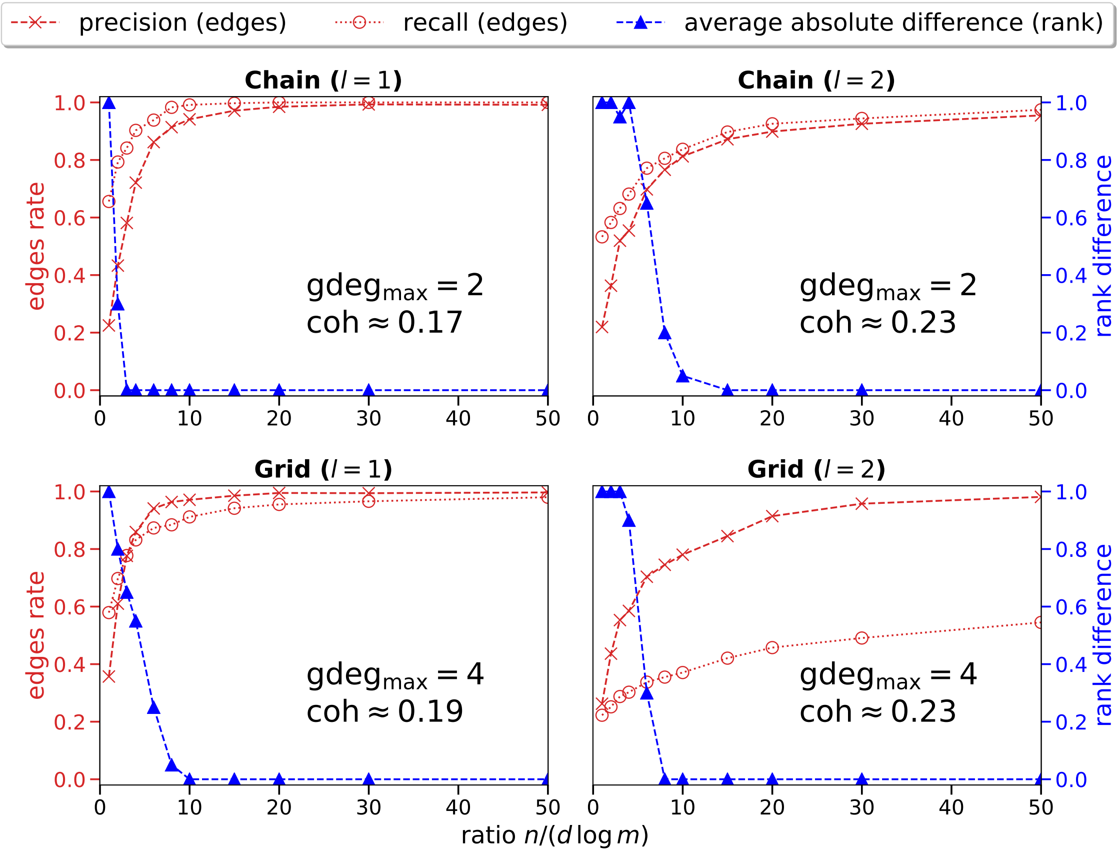

Figure 3.4: Recall, precision, and absolute rank difference averaged over 20 trials for each model and ratio. For each model, the maximum group degree of $\boldsymbol{S}^\star$ and the coherence of $\boldsymbol{L}^\star$ are shown.

3.4.2 Real-world data

We also demonstrate the effectiveness of our fused latent and graphical models on two real-world datasets. The first dataset from the [Open-Source Psychometrics Project] is from a *non-forced* choice vocabulary IQ test (VIQT), where participants can indicate if they do not know an answer, otherwise answers are either correct or wrong. The dataset contains $d = 45$ variables and $n = 12\,173$ samples. The second dataset contains the answers of $n = 165$ test takers to the $d = 72$ questions of the Cambridge face memory test (CFMT) [Itz et al., 2017]. In this dataset, answers with response times below the human reaction time or above some threshold (based on the interquartile range) were assigned to a third category of outliers, otherwise answers are either correct or wrong. Hence, for both datasets, the observed variables are discrete with three possible outcomes.

Figure 3.5 shows estimated fused latent and graphical models for both datasets. The learned models exhibit direct interactions, that is, the answers are not independent given the estimated latent variables. This is in contrast to the common conditional independence assumption in item response theory. Nevertheless, for both models, most observed interactions are explained by a single latent variable. Notably, for the CFMT data, the learned low-rank matrix has a block of positively correlated items in the top left corner. These items correspond to the first block of the CFMT. This block consists of 18 simple questions that most participants get right, hence the correlation.

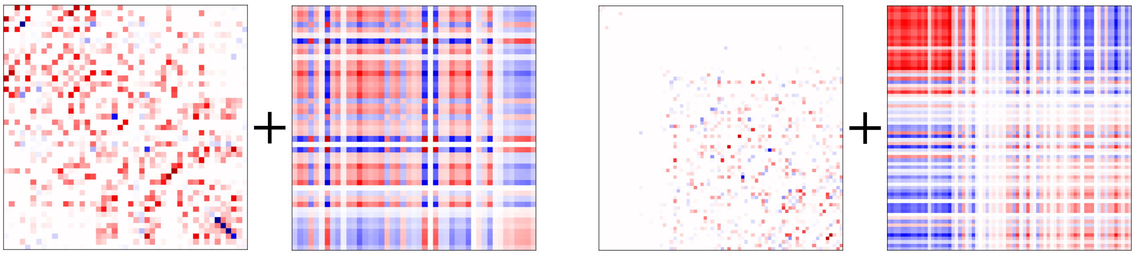

Figure 3.5: Learned decompositions for the VIQT (left) and the CFMT datasets (right). The group-sparse components correspond to direct local dependencies of the observed discrete variables, and the low-rank components represent indirect effects due to the latent continuous variables. Here, red indicates positive and blue indicates negative (conditional) correlations.

3.5 Regularization Parameter Selection

In practice, one fundamental challenge is choosing the regularization parameters λ and γ in Problem (3.7). This is often done using manual, grid, or random search. However, all these methods lack performance guarantees. In this section, we undertake an excursion to a more principled method for selecting regularization parameters. This method is based on Benson's algorithm [Benson, 1998] and was introduced in [Giesen et al., 2019a]. We adapted the method to the semidefinite programming setting in [Giesen et al., 2019b], where we specifically considered fused latent and graphical models in a case study. It should be noted, however, that Benson's algorithm can be used *out of the box* for regularization parameter selection for any other machine learning problem that can be formulated as a regularized convex (semidefinite) program.

In what follows, we provide the details on our extension of Benson's algorithm in Sections 3.5.1-3.5.3. In Section 3.5.4, we present selected experimental results from our work [Giesen et al., 2019b]. Here, it turns out that Benson's algorithm cannot only be used out of the box, but is also efficient and can indeed yield good solutions.

3.5.1 Formulation as a bi-level optimization problem

The objective function of Problem (3.7) can be rescaled to obtain an optimization problem of the form

$$\min_{\boldsymbol{x} \in \mathcal{C}} \quad f_{\boldsymbol{w}}(\boldsymbol{x}) = w_0\, \ell(\boldsymbol{x}) + \sum_{i=1}^{k} w_i\, r_i(\boldsymbol{x}), \tag{$\mathrm{P}_{\boldsymbol{w}}$}$$

where $\mathcal{C} \subseteq V$ is a convex feasible set that is contained in a vector space V over the real numbers, the functions ℓ and r_i, $i = 1, \ldots, k$, are assumed to be convex, and the terms in the objective function $f_{\boldsymbol{w}}$ are weighted by the regularization parameters $\boldsymbol{w} = (w_0, w_1, \ldots, w_k) \in \mathcal{P}^k$, where

$$\mathcal{P}^k = \left\{ \boldsymbol{w} \in \mathbb{R}^{k+1} \,|\, w_i \geq 0 \text{ for } i = 0, \ldots, k \text{ and } \textstyle\sum_{i=0}^{k} w_i = 1 \right\}$$

is the k-dimensional standard simplex. For some given weights $\boldsymbol{w} \in \mathcal{P}^k$, we denote a globally optimal solution to Problem ($\mathrm{P}_{\boldsymbol{w}}$) by $\boldsymbol{x}^{\boldsymbol{w}}$. The goal is to choose good weights, which is often done in the following two-step procedure:

(1) Solve Problem ($\mathrm{P}_{\boldsymbol{w}}$) for 'every' $\boldsymbol{w} \in \mathcal{P}^k$ on *training data.*

(2) Choose $\boldsymbol{w}$ such that some type of *generalization error* $\mathrm{GE}(\boldsymbol{w})$ is minimized on *validation data*, where for the computation of the generalization error the solution $\boldsymbol{x}^{\boldsymbol{w}}$ is used.

These two steps can be jointly understood as the *bi-level optimization problem*

$$\min_{\boldsymbol{w} \in \mathcal{P}^k} \quad \mathrm{GE}(\boldsymbol{w}) \quad \text{s.t.} \quad \boldsymbol{x}^{\boldsymbol{w}} \in \arg\min\,(\mathrm{P}_{\boldsymbol{w}}). \tag{3.15}$$

In this hierarchical optimization problem, the *upper level* strives for a minimal generalization error GE with variable $\boldsymbol{w}$. On the *lower level*, a solution $\boldsymbol{x}^{\boldsymbol{w}}$ of Problem ($\mathrm{P}_{\boldsymbol{w}}$) has to be found for each $\boldsymbol{w}$. In general, Problem (3.15) is *nonconvex*—even if Problem ($\mathrm{P}_{\boldsymbol{w}}$) is convex and GE is a convex objective function [Dempe, 2002]. A practical example is given in Figure 3.8 (see below), where several *local minima* for the generalization error GE exist. In fact, bi-level programming is known to be NP-hard [Hansen et al., 1992].

For solving Problem (3.15), one has to take the set $\{\boldsymbol{x}^{\boldsymbol{w}} : \boldsymbol{w} \in \mathcal{P}^k\}$ of potential candidates into account, that is, the set of solutions of the lower level Problem ($\mathrm{P}_{\boldsymbol{w}}$). In most cases, such as, for fused latent and graphical models, there are no closed-form solutions $\boldsymbol{x}^{\boldsymbol{w}}$. Therefore, we use an approximation that is based on the definition of the *solution gamut* for Problem ($\mathrm{P}_{\boldsymbol{w}}$).

Definition 1 (Solution gamut, [Blechschmidt et al., 2015])**.** Let $\varepsilon > 0$. We call some function $\widehat{x} \colon \mathcal{P}^k \to V$ an *ε-approximative solution gamut* of Problem ($\mathrm{P}_{\boldsymbol{w}}$) if for all $\boldsymbol{w} \in \mathcal{P}^k$

$$\widehat{x}(\boldsymbol{w}) \in \mathcal{C} \quad \text{and} \quad f_{\boldsymbol{w}}(\widehat{x}(\boldsymbol{w})) - f_{\boldsymbol{w}}(\boldsymbol{x}^{\boldsymbol{w}}) \leq \varepsilon. \qquad \blacktriangle$$

Note that the *full solution gamut* given by the set $\{\boldsymbol{x}^{\boldsymbol{w}} : \boldsymbol{w} \in \mathcal{P}^k\}$ corresponds to the ε-approximative solution gamut with $\varepsilon = 0$.

In the following, we will present a variant of *Benson's dual algorithm*, which is an established method from the area of *vector optimization.* It can be used to compute an ε-approximative solution gamut of Problem ($\mathrm{P}_{\boldsymbol{w}}$). Since the algorithm (if it converges) yields a finite representation of the ε-approximative solution gamut $\widehat{x}$, minimization of the upper level objective function GE just boils down to function evaluations.

3.5.2 Basics of vector optimization

For applying Benson's algorithm, we study Problem ($\mathrm{P}_{\boldsymbol{w}}$) in the context of vector optimization. For that, consider the following problem

$$\min_{\boldsymbol{x} \in \mathcal{C}} \; F(\boldsymbol{x}) = \Big(\ell(\boldsymbol{x}), r_1(\boldsymbol{x}), \ldots, r_k(\boldsymbol{x})\Big). \qquad (P_{\text{vec}})$$

Here, the objective function $F \colon V \to \mathbb{R}^{k+1}$ is vector-valued and minimized w.r.t. the component-wise partial ordering $\leq_{\mathbb{R}^{k+1}_+}$ on $\mathbb{R}^{k+1}$. It is given by

$$\boldsymbol{y}_1 \leq_{\mathbb{R}^{k+1}_+} \boldsymbol{y}_2 \quad \Longleftrightarrow \quad \boldsymbol{y}_2 - \boldsymbol{y}_1 \in \mathbb{R}^{k+1}_+ = \{\boldsymbol{y} \in \mathbb{R}^{k+1} \mid y_i \geq 0,\, i = 1, \ldots, k+1\}.$$

Note that Problem ($\mathrm{P}_{\boldsymbol{w}}$) is the *weighted sum scalarization* of Problem (P_{vec}) with objective function $\boldsymbol{w}^\mathsf{T} F(\boldsymbol{x})$. Let us now define the minimizers of Problem (P_{vec}):

Definition 2. A point $\boldsymbol{x}^* \in \mathcal{C}$ is called a *weak minimizer* of Problem (P_{vec}) if

$$(\{F(\boldsymbol{x}^*)\} - \operatorname{int} \mathbb{R}^{k+1}_+) \cap F(\mathcal{C}) = \varnothing,$$

where $\operatorname{int} \mathbb{R}^{k+1}_+ = \{\boldsymbol{y} \in \mathbb{R}^{k+1} \mid y_i > 0,\, i = 1, \ldots, k+1\}$ is the interior of $\mathbb{R}^{k+1}_+$, and $F(\mathcal{C}) = \{F(\boldsymbol{x}) \mid \boldsymbol{x} \in \mathcal{C}\} \subseteq \mathbb{R}^{k+1}$ is the image of the feasible set. ▲

Some weak minimizers are visualized in Figure 3.6. Giesen et al. [2019a] observed that the *full solution gamut* (Definition 1 with $\varepsilon = 0$) coincides with the set of *weak minimizers* of Problem (P_{vec}). The image of the set of all weak minimizers is called the *Pareto set* or *Pareto frontier*.

In general, there are infinitely many weak minimizers. However, without closed-form solutions, we can compute only a finite number of weak minimizers. This finite number of minimizers should approximate the set of weak minimizers well. Hence, it is important to understand the geometry of Problem (P_{vec}), in particular the *upper image*

$$\mathcal{U} = \text{closure}(F(\mathcal{C}) + \mathbb{R}_+^{k+1}).$$

A non-empty finite set $M \subseteq \mathcal{C}$ of weak minimizers yields an inner polyhedral approximation of the upper image given by

$$\text{conv}\, F(M) + \mathbb{R}_+^{k+1}.$$

The following definition can be used to assess how close such an approximation is to the upper image.

Definition 3. Let $\boldsymbol{c} \in \text{int}\, \mathbb{R}_+^{k+1}$ be an arbitrary but fixed direction. Assume that Problem (P_{vec}) is bounded in the sense that it holds $\mathcal{U} \subseteq \{\boldsymbol{y}\} + \mathbb{R}_+^{k+1}$ for some $\boldsymbol{y} \in \mathbb{R}^{k+1}$. Then, a non-empty finite set $M \subseteq \mathcal{C}$ of weak minimizers is called a *(weak) ε-solution* to Problem (P_{vec}) if

$$\mathcal{U} \subseteq \text{conv}\, F(M) + \mathbb{R}_+^{k+1} - \varepsilon \boldsymbol{c}. \qquad \blacktriangle$$

Note that a weak ε-solution is only defined w.r.t. the direction $\boldsymbol{c}$. For a weak ε-solution, we denote the inner polyhedral approximation of the upper image $\mathcal{U}$ obtained from M as $\mathcal{I}_\varepsilon = \mathcal{I}_\varepsilon(M) = \text{conv}\, F(M) + \mathbb{R}_+^{k+1}$, see Figure 3.6. Since M is a weak ε-solution, no point of $\mathcal{U}$ has a distance larger than ε in direction $\boldsymbol{c}$ to $\mathcal{I}_\varepsilon$. Hence, a weak ε-solution also yields an *outer* polyhedral approximation of $\mathcal{U}$ given by $\mathcal{I}_\varepsilon - \varepsilon \boldsymbol{c}$.

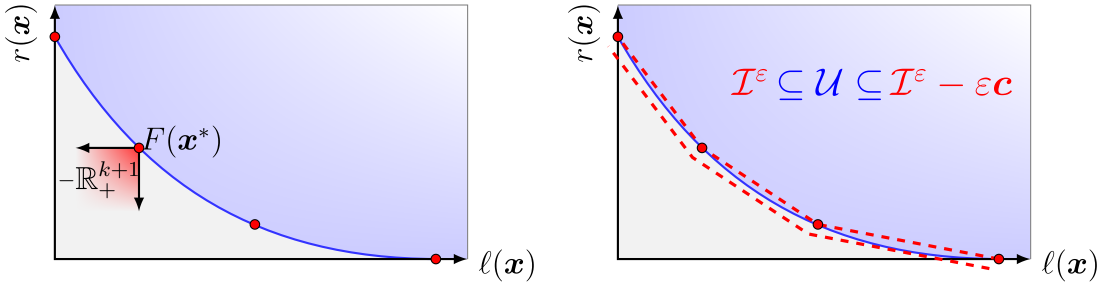

Figure 3.6: *Left:* Pareto frontier (blue line) with weak minimizers of Problem (P_{vec}) (red points), see Definition 2. *Right:* Upper image $\mathcal{U}$ with an inner polyhedral approximation $\mathcal{I}^\varepsilon$ and an outer polyhedral approximation $\mathcal{I}_\varepsilon - \varepsilon \boldsymbol{c}$. Both were obtained from a weak ε-solution, see Definition 3.

3.5.3 Adaptive Benson algorithm

We want to compute weak ε-solutions for Problem (P_{vec}). For this, we develop an *adaptive* variant of Benson's dual algorithm. The class of dual Benson algorithms is designed to approximate the upper image $\mathcal{U}$. It proceeds by iteratively generating a growing sequence of inner polyhedral approximations until an ε-solution for a prescribed accuracy $\varepsilon > 0$ is obtained. Details on the class of Benson algorithms can, for example, be found in [Benson, 1998; Giesen et al., 2019a].

Polyhedra are crucial for Benson's algorithm. Each non-empty convex polyhedral set $A \subseteq \mathbb{R}^{k+1}$ is either given in *H-representation*, that is, as the intersection of finitely many half spaces

$$A = \bigcap_{i=1}^{r} \{\boldsymbol{y} \in \mathbb{R}^{k+1} \mid (\boldsymbol{w}^i)^\mathsf{T}\boldsymbol{y} \geq b_i\} \quad \text{for } \boldsymbol{0} \neq \boldsymbol{w}^i \in \mathbb{R}^{k+1}, b_i \in \mathbb{R}, \quad i = 1, \dots, r,$$

or in *V-representation*, that is, as a set of vertices and directions

$$A = \operatorname{conv}\{\boldsymbol{v}_1, \dots, \boldsymbol{v}_s\} + \operatorname{cone}\{\boldsymbol{d}^1, \dots, \boldsymbol{d}^t\}, \qquad s, t \in \mathbb{N}, s \neq 0,$$

where $\boldsymbol{v}_i \in \mathbb{R}^{k+1}$ are the vertices and $\boldsymbol{0} \neq \boldsymbol{d}^j \in \mathbb{R}^{k+1}$ are the directions. The conversion between both representations is respectively done by *vertex* and *facet enumeration* [Bremner et al., 1998]. For Algorithm 1 and below, the cone for the V-representation is always $\mathbb{R}^{k+1}_+$.

Algorithm 1 Adaptive Dual Benson Algorithm

Input: Problem data (F, $\mathcal{C}$), initialization $\mathcal{V}$, set of checked hyperplanes/normals $T = \{\boldsymbol{w}_0\}$, direction $\boldsymbol{c}$, initial accuracy $\varepsilon = \varepsilon_0$
Output: nodes of V-representation $\mathcal{V}$

```
function AdaptiveBensonAlgorithm
    repeat
        $\mathcal{V}_\varepsilon \leftarrow \{F(\boldsymbol{x}^{\boldsymbol{w}}) \in \mathcal{V} \mid d_{\boldsymbol{c}}(\boldsymbol{x}^{\boldsymbol{w}}) \geq \varepsilon\}$
        compute H-representation $\mathcal{I}_\varepsilon$ of V-representation $\operatorname{conv}(\mathcal{V}_\varepsilon) + \mathbb{R}^{k+1}_+$
        repeat
            choose $\boldsymbol{w} \in \mathcal{I}_\varepsilon \setminus T$
            $\boldsymbol{x}^{\boldsymbol{w}} \leftarrow \arg\min\,(\mathrm{P}_{\boldsymbol{w}})$
            if $d_{\boldsymbol{c}}(\boldsymbol{x}^{\boldsymbol{w}}) > \varepsilon$ then
                $\mathcal{V}_\varepsilon \leftarrow \mathcal{V}_\varepsilon \cup \{F(\boldsymbol{x}^{\boldsymbol{w}})\}$
                update H-representation $\mathcal{I}_\varepsilon$ of V-representation $\operatorname{conv}(\mathcal{V}_\varepsilon) + \mathbb{R}^{k+1}_+$
            end if
            $T \leftarrow T \cup \{\boldsymbol{w}\}$
        until $\mathcal{I}_\varepsilon \setminus T = \varnothing$
        $\mathcal{V} \leftarrow \mathcal{V}_\varepsilon \cup \mathcal{V}$
        $\varepsilon \leftarrow \varepsilon/2$
    until Stopping Criterion
end function
```

For the standard Benson algorithm, an approximation accuracy ε must be chosen beforehand. However, in practice, ε cannot be chosen generically. Therefore, we propose Algorithm 1 as an adaptive variant of Benson's dual algorithm. It starts by calculating an initial ε_0-approximation $\mathcal{I}_{\varepsilon_0}$ of the upper image $\mathcal{U}$ for a coarse accuracy $\varepsilon_0 > 0$. Then, the accuracy is successively refined by setting $\varepsilon_{i+1} = \varepsilon_i/2$ until the approximation of $\mathcal{U}$ satisfies some stopping criterion. This leads to a growing sequence of inner polyhedral approximations given by

$$\mathcal{I}_{\varepsilon_0} \subseteq \mathcal{I}_{\varepsilon_1} \subseteq \ldots \subseteq \mathcal{I}_{\varepsilon_i} \subseteq \ldots \subseteq \mathcal{U}.$$

Here, an (intermediate) inner approximation $\mathcal{I}_{\varepsilon_i}$ is calculated as follows: First, we choose a hyperplane $\boldsymbol{w}^\mathsf{T}\boldsymbol{x} = b$ from the H-representation of the current inner approximation. Then, we move this hyperplane outwards until it becomes a supporting hyperplane of the upper image $\mathcal{U}$, see Figure 3.7. The supporting hyperplane contacts $\mathcal{U}$ in the point $F(\boldsymbol{x}^{\boldsymbol{w}})$, where $\boldsymbol{x}^{\boldsymbol{w}}$ is the solution to the scalarized Problem $(\mathrm{P}_{\boldsymbol{w}})$.

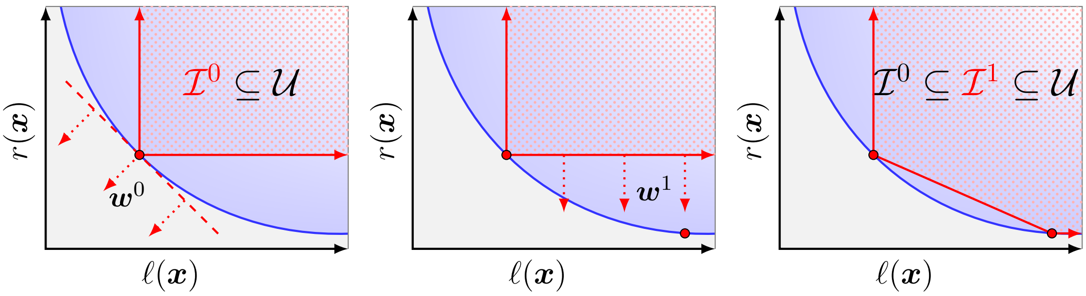

Figure 3.7: *Left:* Initialization of Algorithm 1 with the solution to Problem $(\mathrm{P}_{\boldsymbol{w}})$ with weights $\boldsymbol{w} = \boldsymbol{w}^0$. The dashed line represents the hyperplane that contacts the upper image $\mathcal{U}$ in the point $F(\boldsymbol{x}^{\boldsymbol{w}_0})$. *Middle and right:* First iteration of Algorithm 1 for the hyperplane with normal $\boldsymbol{w}^1$. The hyperplane is moved outwards until it becomes a supporting hyperplane of the upper image, where the weak minimizer $F(\boldsymbol{x}^{\boldsymbol{w}_1})$ is a contact point.

We want to compute an ε_i-solution w.r.t. a fixed direction $\boldsymbol{c}$. Therefore, we only need to refine the current inner polyhedral approximation around the vertex $F(\boldsymbol{x}^{\boldsymbol{w}})$ if the hyperplane $\boldsymbol{w}^\mathsf{T}\boldsymbol{x} = b$ was moved outwards at least distance ε_i in direction $\boldsymbol{c}$, where the distance is given as

$$d_{\boldsymbol{c}}(\boldsymbol{x}^{\boldsymbol{w}}) = \cos(\boldsymbol{c}, \boldsymbol{w})\left(b - \boldsymbol{w}^\mathsf{T} F(\boldsymbol{x}^{\boldsymbol{w}})\right) = \frac{\boldsymbol{c}^\mathsf{T}\boldsymbol{w}}{\|\boldsymbol{c}\|_2\,\|\boldsymbol{w}\|_2}\left(b - \boldsymbol{w}^\mathsf{T} F(\boldsymbol{x}^{\boldsymbol{w}})\right).$$

Hence, if $d_{\boldsymbol{c}}(\boldsymbol{x}^{\boldsymbol{w}}) \geq \varepsilon_i$, we add the vertex $F(\boldsymbol{x}^{\boldsymbol{w}})$ to the current V-representation and then we update the H-representation by facet enumeration. If $d_{\boldsymbol{c}}(\boldsymbol{x}^{\boldsymbol{w}}) < \varepsilon_i$, we continue and check the next hyperplane. The ε_i-approximation of $\mathcal{U}$ is completed when no unchecked hyperplanes remain in the H-representation.

For the next iteration of Algorithm 1 with accuracy $\varepsilon_{i+1} = \varepsilon_i/2$, the V-representation is initialized with all vertices $F(\boldsymbol{x}^{\boldsymbol{w}})$ where $d_{\boldsymbol{c}}(\boldsymbol{x}^{\boldsymbol{w}}) \geq \varepsilon_{i+1}$. In the neighborhood of these vertices the inner polyhedral approximation may need further improvement.

The other vertices are kept and may be used in a later iteration of Algorithm 1 (moreover, the corresponding solutions may have good generalization errors).

Algorithm 1 can be initialized with a V-representation that has a single node $F(\boldsymbol{x}^{\boldsymbol{w}_0})$ for some initial weight vector $\boldsymbol{w}_0 \in \mathbb{R}^{k+1}$. For instance, one can choose $\boldsymbol{w}_0 = (1, \ldots, 1)/\sqrt{k+1}$ and set the initial nodes of the V-representation to $\mathcal{V} = \{F(\boldsymbol{x}^{\boldsymbol{w}_0})\}$. Next, as stopping criteria for Algorithm 1 one of the following can be used: (i) a target ε-approximation of $\mathcal{U}$ is attained, (ii) a maximum number of iterations has been completed, or (iii) the generalization error has improved by a certain degree. Note that generalization errors can be computed on the fly by simple function evaluations.

Convergence and complexity. Löhne et al. [2014] analyzed the case, where the feasible set $\mathcal{C}$ for Problem (P_{vec}) is exclusively described by polyhedral cone constraints. They showed that if Benson's (dual) algorithm terminates, then it also works correctly, that is, it returns a weak ε-solution. Their result can be extended to semidefinite constraints since by [Löhne et al., 2014, Remark 3 (Section 4.3)] it is only required that (i) $\operatorname{int} \mathcal{C} \neq \varnothing$ (Slater's condition) and (ii) Problem ($\mathrm{P}_{\boldsymbol{w}}$) has a solution for all $\boldsymbol{w} \in \mathcal{P}^k$. This is the case in our setting.

A lower bound for the number of optimization problems that must be solved for obtaining an ε-approximative solution gamut is $\Omega(\varepsilon^{-k/2})$ [Blechschmidt et al., 2015]. Theorems 3 and 4 in [Kamenev, 1994] also give the general upper bound $\mathcal{O}(\varepsilon^{-k})$ for Benson-type algorithms, and they give the sharp upper bound $\mathcal{O}(\varepsilon^{-k/2})$ provided that the upper image $\mathcal{U}$ has a twice continuously differentiable boundary. These bounds have not yet been transferred to the exact setting of Algorithm 1. We investigate the complexity in our experiments.

3.5.4 Experiments

In [Giesen et al., 2019b], the goal was to compare the performance of Algorithm 1 against the *solution gamut method* from [Blechschmidt et al., 2015], which is the only other known method to obtain ε-approximative solution gamuts. Here, we do not show the results of the comparison since the main focus of this thesis are fused latent and graphical models. Instead, we restrict ourselves to selected experimental results from [Giesen et al., 2019b] that demonstrate the effectiveness of using Algorithm 1 for fused latent and graphical models. We study two aspects: the quality of the solutions (in terms of the generalization error GE, here the negative log-likelihood function value) and the computational efficiency.

Setup and preparation. For solving the optimization problems, we use the ADMM-based algorithm discussed in [Ma et al., 2013], see also Section 3.3, and we use CDD [Bremner et al., 1998] for facet enumeration. All experiments were run on a Linux machine with an Intel Core i5-2500K (4 × 3.30 GHz) CPU and 16 GB RAM.

In the experiments we use the following data sets [Tsanas et al., 2014; Higuera et al., 2015; Zhou et al., 2014; Dua and Karra Taniskidou, 2017]

- `GENE1` with $n = 100$ features and $m = 255$ samples,
- `TCGA` with $n = 500$ features and $m = 801$ samples,
- `MICE` with $n = 81$ features and $m = 552$ samples,
- `ROSETTA` with $n = 100$ features and $m = 301$ samples,
- `SONAR` with $n = 60$ features and $m = 208$ samples,
- `OR70` with $n = 70$ features and $m = 1059$ samples,
- `LSVT` with $n = 310$ features and $m = 126$ samples,
- `S&P500` with $n = 471$ features and $m = 60$ samples.

These data sets are from different applied areas: `GENE1`, `TCGA`, and `MICE` are biological data sets, `ROSETTA` and `SONAR` are geological data sets, `OR70` was recorded for investigating the geographical origins of music, `LSVT` is about voice rehabilitation in psychology, and `S&P500` includes monthly stock return data from major US companies over the course of 5 years. From `S&P500` we removed 29 companies because their data was incomplete. From the original data sets `GENE1`, `ROSETTA`, and `TCGA` we selected the n features with the highest variance, similarly as Chandrasekaran et al. [2012] who also used only subsets of `GENE1` and `ROSETTA`. Note that all data sets have only continuous features since the paper [Giesen et al., 2019b] considered only this setting. Data sets were split randomly into training and validation data in a 2:1 ratio. We also centralized and standardized the data using empirical means and standard deviations of the training data. The generalization error, here the negative log-likelihood function value, is computed on the validation data.

For the experiments, we reformulated Problem (3.7) such that it conforms to the theory in Section 3.5.1, where weights are chosen from a standard simplex. More specifically, for $\alpha, \beta \geq 0$ with $\alpha + \beta \leq 1$, we used the weight α for the ℓ_1-norm, β for the trace (nuclear) norm, and $(1 - \alpha - \beta)$ for the negative log-likelihood term. We always use the fixed direction $\boldsymbol{c} = (1, 1, 1)^\mathsf{T}$ for the computation of ε-solutions in Algorithm 1.

Solution quality. Here, we only show the results from the `LSVT` data set [Tsanas et al., 2014] after stopping Algorithm 1 at the accuracy of $\varepsilon = 2^6$ (with starting $\varepsilon_0 = 2^{10}$). As a baseline for the evaluation of the solution quality, we performed a grid search on a fine grid with more than 5 000 points. The generalization errors for the solutions from the grid points are shown in Figure 3.8. Moreover, Figure 3.8 shows that the accuracy $\varepsilon = 2^6$ is already sufficient for finding good solutions with different algebraic properties, that is, different sparsity and rank. Although the search region for the upper level problem, which minimizes the GE, is non-convex, Benson's algorithm found solutions close to all local minima. Hence, in practice the best solutions returned by Benson's algorithm should be taken into account because

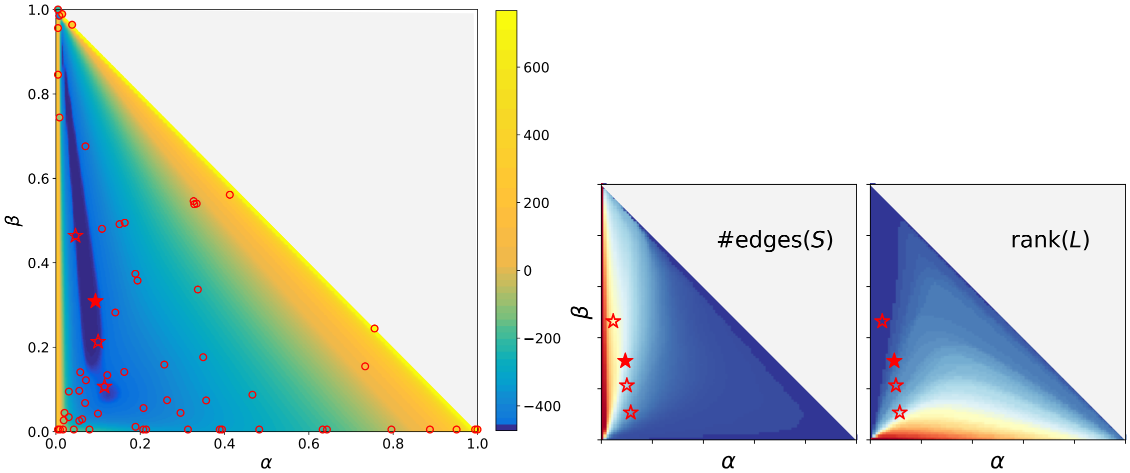

Figure 3.8: The left figure shows the generalization errors from the baseline grid solutions for the `LSVT` data set. Their corresponding sparsity and rank patterns are shown in the two plots on the right, where cold colors indicate high sparsity (left) and low rank (right). Solutions from running Algorithm 1 are marked by circles. The four best solutions are highlighted by stars. The best solution has a filled star. Its corresponding sparse and low-rank decomposition is shown on the left in Figure 3.9. The other decomposition in Figure 3.9 belongs to the Benson solution near the smaller region with a local minimum.

they may provide alternatives in terms of their algebraic properties, see Figure 3.9 for an example.

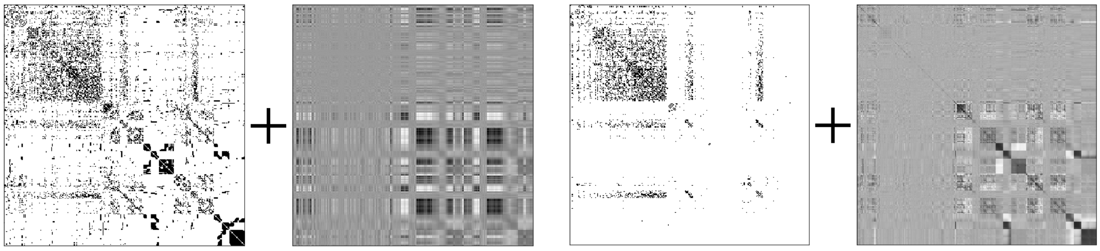

Figure 3.9: Two decompositions of learned fused latent and graphical models for the `LSVT` data. The left decomposition corresponds to the solution that is optimal in terms of the generalization error (compare Figure 3.8).

Computational efficiency. The pessimistic known theoretical upper bound for Benson's algorithm is only in $O(1/\varepsilon^2)$, while the sharp bound $\Omega(1/\varepsilon)$ holds only under certain regularity assumption. We performed experiments to see which scenario is more realistic in practice. For that, for each data set, we counted the number of iterations of Algorithm 1 until selected ε-approximations of the upper image (ε-solutions) were achieved. Figure 3.10 shows the resulting log-log complexity plot. It turns out that the adaptive Benson algorithm experimentally matches the sharp bound $\Omega(1/\varepsilon)$. This suggests that the optimistic upper bound for Benson's algorithm is realistic for fused latent and graphical models. However, we would like to point out that in practice, there also incurs some overhead for facet enumerations in the adaptive Benson algorithm. This should be considered when deciding which method to use for regularization parameter selection.

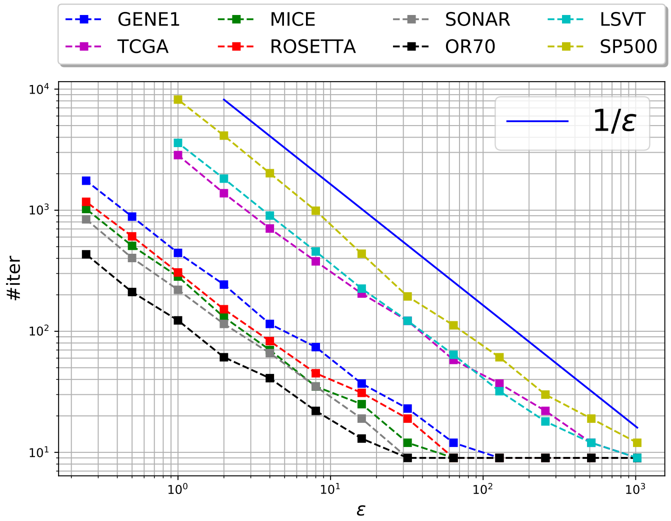

Figure 3.10: Log-log complexity plot for the adaptive Benson algorithm (Algorithm 1): The y-axis shows the number of iterations that were required for obtaining selected ε-solutions for the respective data sets. Here, an iteration consists of checking one hyperplane.

3.6 Concluding Remarks

In this chapter, we investigated fused latent and graphical models for mixed observed discrete *and* quantitative, conditional Gaussian variables. These models are characterized by a group-sparse + low-rank decomposition of the pairwise interaction parameter matrix. We have shown that learning such models using the convex optimization Problem (3.7) can produce consistent estimates in the high-dimensional setting. Consistent recovery is possible under certain assumptions that we motivated carefully. The assumptions mostly ensure two important prerequisites: First, that the observations have sufficient quality, and second, that the group-sparse + low-rank matrix decomposition is identifiable.

To practically estimate fused latent and graphical models, we make use of the pseudo-likelihood, which does not require the computation of costly normalization constants. Moreover, it is often more numerically stable. We implemented a proximal-gradient based ADMM solver for Problem (3.7), where on default the pseudo-likelihood is used. The solver is part of the Python code repository for model selection of conditional Gaussian distributions that we published under `https://github.com/franknu/cgmodsel`.

In this chapter, we conducted several experiments to support the theory. First, we verified experimentally that consistent recovery becomes easier if there are not too many non-zero groups per row/column of the group-sparse matrix and if the low-rank matrix is spread-out. Second, we learned fused latent and graphical models from real-world data, demonstrating that observed data from item response theory studies can be conditionally dependent given the latent variables—in contrast to the common assumption. This shows that modeling direct interactions via fused latent and graphical models is reasonable. Apart from item response theory models with only discrete variables, applications with mixed observed discrete and quantitative data can be found in many different domains, for instance, in medicine [Sammel et al., 1997], in biology in form of metabolic networks or gene expression data, or in economics, such as, census data [Lee and Hastie, 2015].

Finally in this chapter, we introduced a method for selecting the regularization parameters of Problem (3.7) in an efficient and principled way. This method is based on Benson's algorithm. It approximates the set of all possible solutions to Problem (3.7) that can be obtained from using different regularization parameters. From among the finite number of solutions that are computed during the approximation, the best is chosen. Let us now discuss some potential future research directions.

Future directions. Though the class of distributions that we considered in this chapter is quite general, further generalizations are possible. For example, one can allow higher-order interactions between the variables. Cheng et al. [2017] already studied sparse graphical models for CG distributions with triple interactions, where they allowed a dependence of the precision matrix (quantitative-quantitative interactions) on the discrete variables. In their most general form, the *canonical representation* of a CG distribution on $\mathcal{X} \times \mathcal{Y} = \prod_{i=1}^{d} \{0, \dots, m_i\} \times \mathbb{R}^q$ is given by

$$p(\boldsymbol{x}, \boldsymbol{y}) \propto \exp\left(q(\boldsymbol{x}) + \nu(\boldsymbol{x})^\intercal \boldsymbol{y} - \frac{1}{2}\boldsymbol{y}^\intercal \Lambda(\boldsymbol{x})\boldsymbol{y}\right), \qquad (\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X} \times \mathcal{Y}, \tag{3.16}$$

where for all $\boldsymbol{x} \in \mathcal{X}$ we have $\mathbf{0} \prec \Lambda(\boldsymbol{x}) \in \mathrm{Sym}(q)$, $q(\boldsymbol{x})$ defines arbitrary interactions between the discrete variables, and $\nu(\boldsymbol{x}) \in \mathbb{R}^q$ defines the interactions between discrete and quantitative variables. Similarly as in Appendix C.2, it can be shown that marginalizing out some of the quantitative, conditional Gaussian variables in (3.16) yields a marginal model $p(\boldsymbol{x}, \boldsymbol{y}')$ whose conditional (Gaussian) distributions $p(\boldsymbol{y}' \mid \boldsymbol{x})$ are still characterized by a sparse + low-rank decomposition of their respective precision matrices. Hence, for each discrete outcome, there is a sparse + low-rank decomposition. Suitable learning methods for estimating such collections of sparse + low-rank decompositions would yet need to be devised. It is not known if the approach by using regularized convex likelihood problems remains tractable. Moreover, it is not clear if the consistency guarantees can be generalized in some sense. Because of the significantly larger number of parameters, there will probably be additional challenges. Nevertheless, the author believes that it should be possible to obtain similar results as for the pairwise fused latent and graphical models from this chapter. Note that for practical learning of CG models with higher-order interactions one would probably need to resort to pseudo-likelihood estimation.

For pseudo-likelihood estimation, it would be interesting to also obtain learning guarantees for estimating fused latent and graphical models. Chen et al. [2016] showed consistency for the pseudo-likelihood estimation of binary fused latent and graphical models. However, their consistency result is much weaker than the results that follow from Theorem 3.3 in the high-dimensional setting. Instead, the result by

Chen et al. [2016] only states convergence of the parameters in probability, without providing convergence rates. While their result is a good justification for using pseudo-likelihood in practice, it would be interesting to see if similar guarantees in the high-dimensional setting can be obtained for pseudo-likelihood estimation of fused latent and graphical models.

As a final thought, one might also consider higher-order interaction models and try to decompose the parameters in sparse and low-rank parts. In doing so, the latent-variable interpretation will likely be lost in most cases. However, composing the presumably large number of parameters in higher-order models can help to balance model complexity. Since the parameters of higher-order interaction models are tensors, the regularization techniques in the convex optimization approach must be generalized to tensors. There are several possibilities, for example, using generalizations of the nuclear norm to tensors [Tomioka and Suzuki, 2013; Zhang et al., 2014].

Chapter 4

Conclusion

In this thesis, we studied robust principal component analysis and fused latent and graphical models. All models have in common that they feature low-rank and group-sparse decompositions of matrices. For robust principal component analysis, a corrupted data matrix is decomposed directly, where the low-rank component represents the principal components, and the group-sparse component accounts for the data corruption. In contrast, for fused latent and graphical models, the matrix of pairwise interaction parameters for the observed variables is decomposed. Here, the group-sparse component corresponds to direct interactions among the observed variables, and the low-rank component describes indirect interactions that can be attributed to a presumably small number of latent quantitative variables.

We showed that low-rank and group-sparse matrix decompositions can be learned efficiently via convex regularized optimization problems. Here, nuclear norm regularization is used to induce low rank, and $\ell_{1,2}$-norm regularization is used to promote group sparsity on a matrix. For fused latent and graphical models, there is an additional likelihood term in the objective function. This term ensures a good fit of the learned probabilistic model to the data. In all optimization problems, the different terms of the objective function are weighted with regularization parameters that must be chosen beforehand.

As a central contribution of this thesis, we showed that the estimation of models via convex optimization comes with strong guarantees: In many cases, it is possible to recover an assumed true decomposition exactly (robust principal component analysis) or consistently in the high-dimensional setting (fused latent and graphical models). An important necessary condition that facilitates recovery is the identifiability of the low-rank and group-sparse matrix decomposition. Here, a decomposition is identifiable if the components cannot be confused. This is the case if neither component is low rank and group sparse at the same time. We showed that identifiability requires the respective tangent spaces to the low-rank and group-sparse matrix varieties to be transverse, which means that their intersection only contains the origin (zero vector). Moreover, our theoretical results predict that exact and consistent recovery is respectively possible for a range of regularization parameters.

We experimentally verified the theoretical results on synthetic data, where the true underlying model is known. For robust principal component analysis, we showed that successful recovery of a random decomposition is more likely if the rank and the number of non-zero groups of the respective components are not too large. To determine if exact recovery is possible, we explicitly searched for the predicted range of regularization parameters for which the convex learning problem yields the correct solution. Next, for fused latent and graphical models, we tested the influence of maximum group degree and incoherence on recovery rates. If these quantities are small, then the components can be recovered more easily because it is harder to confuse them. We also performed a variety of experiments on real-world data to demonstrate the usefulness of the models.

Important previous works for this thesis are Candès et al. [2011]; Chandrasekaran et al. [2011, 2012]. In many aspects, these works laid the foundations for the models and techniques used throughout this thesis. However, they considered only regular sparsity and thus problems with ℓ_1-norm regularization. An important contribution of this thesis was to extend the original models and their respective theoretical analyses to general settings that involve group sparsity: First, for robust principal component analysis, previously only corruptions of single entries and whole data points were considered. Our model allows more general data corruption mechanisms that affect groups of measurements. Second, we generalized fused latent and graphical models to distributions with new types of observed variables, namely conditional Gaussian distributions with observed discrete and quantitative variables.

Most contributions from this thesis are also published in separate works [Nussbaum and Giesen, 2019a, 2020a,b, 2021]. For fused latent and graphical models, we also undertook an excursion by adapting Benson's algorithm to the problem of selecting suitable regularization parameters [Giesen et al., 2019b]. The proposed variant of the algorithm can be used out of the box for large classes of optimization problems. It poses a principled alternative to basic methods, such as, grid or random search.

Future directions. In addition to the directions that we outlined in the respective chapters, we would like to point out the following directions for future research.

Applications. In this thesis, we demonstrated some real-world applications. We did so mostly on a qualitative level. Deploying the models in real applications can be the subject of future research. For example, it could be interesting to make use of discovered direct interactions among test items from psychometric tests. Classical item response theory models commonly assume that these direct dependencies should not exist. Therefore, fused latent and graphical models may even be helpful for improving psychometric tests. Chen et al. [2018] already sparked research in this direction, but the methods yet need to develop before becoming a standard.

Large-scale problems. The ADMM-based solvers for the optimization problems in this thesis work well for moderate problem sizes. However, there are clear limitations when the problem dimensions become large. The main limitation of the ADMM algorithm for robust PCA in Section 2.3 lies in the cost for computing a partial singular value decomposition. We used an inexact method (randomized singular value thresholding [Halko et al., 2011]). This method can be efficient, but the convergence guarantee of ADMM is lost. Our experiments indicate that convergence is not harmed when only a few singular values are required. However, additional testing and numerical experiments are required to analyze the correctness of the solver, especially for larger problems. Next, for learning fused latent and graphical models, some limitations were already discussed in Section 3.3.2: The proposed proximal-gradient based ADMM algorithm calls an iterative optimization algorithm inside its outer loop. It does so to compute the proximal mapping of the pseudo-likelihood, which does not permit closed-form solutions. The nested iterations are inefficient and responsible for the fact that the current solver scales only up to several hundreds of variables, perhaps a few thousands. A more efficient solver might have to avoid the costly computation of the proximal mapping of the pseudo-likelihood. This can be the subject of future research.

Model generalizations. The models from this thesis can be generalized further. We outlined extensions of robust principle component analysis to the more general data corruption mechanism that affects *sub-matrices* of the data matrix. Also a weighted $\ell_{1,2}$-norm could be used to encode different prior beliefs about the probabilities that groups of measurements are corrupted. For fused latent and graphical models, generalizations to higher-order interactions could be the subject of future research. Another direction could be to go beyond conditional Gaussian distributions by allowing even more general types of observed variables. Particularly interesting are exponential family distributions, for example, based on the Poisson or the exponential distribution. However, it can already be challenging to define multivariate generalizations of these univariate distributions [Yang et al., 2013, 2015; Inouye et al., 2016]. An important restriction of fused latent and graphical models is that all interactions are purely associative, that is, the effect of the latent variables cannot be interpreted causally. It would be interesting to rigorously establish connections between fused latent and graphical models and causality. A final promising extension can be models that, in addition to quantitative latent variables, allow for discrete latent variables. Discrete latent variables entail mixture distributions on the observed variables, which can improve the expressive power of the models.

Each of the suggested generalizations can advance probabilistic modeling and thereby, the tools that we have for understanding the world. Ultimately, we strive for models that (a) advance our scientific understanding by revealing fundamental truths, and (b), provide guidance for complex decisions. Better models also tend to be more useful. To achieve the aforementioned goals, it is crucial that the models capture essential aspects of our environment and its underlying processes. As there is always room for improvement, this thesis represents only a small step. However, every journey consists of small steps. After all, small steps are to be preferred over too complex and uncontrollable ones. This is much in analogy to the centuries-old problem-solving principle that endorses simplicity over complexity: Ockham's razor.

Appendix A

Notation

A.1 Notational Conventions

Expression	Notation
vectors	bold lower case: $\boldsymbol{e}, \boldsymbol{v}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}$
matrices	bold upper case: $\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{X}, \boldsymbol{M}, \boldsymbol{N}, \boldsymbol{A}$, also $\boldsymbol{\Delta}, \boldsymbol{\Theta}, \boldsymbol{\Gamma}, \boldsymbol{\Lambda}, \boldsymbol{\Sigma}$
matrix spaces	calligraphic letters: $\mathcal{Q}, \mathcal{T}, \mathcal{J}$
scalars	indices: i, j, k, l, s, t
	regularization parameters: λ, γ, α
	group structure: m_1, m_2, ..., η
	number of sample points: n
	number of variables/features: d, m, q, w
	'problem-specific' constants: β, δ, α, ..., C, c_0, c_1, ..., C_1, C_2, ...

A.2 Specific Notation

Notation	Definition
$\|\boldsymbol{X}\|$	spectral norm
$\|\boldsymbol{X}\|_*$	nuclear norm (for $\boldsymbol{X} \succeq 0$ it holds $\|\boldsymbol{X}\|_* = \mathrm{tr}(\boldsymbol{X})$)
$\|\boldsymbol{X}\|_{1,2}$	$\ell_{1,2}$-group norm
$\|\boldsymbol{X}\|_{\infty,2}$	$\ell_{\infty,2}$-group norm
$\|\boldsymbol{X}\|_F$	Frobenius norm
$\|(\boldsymbol{M}, \boldsymbol{N})\|_\gamma$	γ-norm, $\|(\boldsymbol{M}, \boldsymbol{N})\|_\gamma = \max\{\|\boldsymbol{M}\|_{\infty,2}/\gamma, \|\boldsymbol{N}\|\}$
$\partial\|\boldsymbol{X}\|$	norm sub-differential at point $\boldsymbol{X}$
$\mathrm{Sym}(d)$	the set of symmetric $(d \times d)$-matrices
$\mathcal{S}(s)$	group-sparse matrix variety (matrices with as most s non-zero groups)
$\mathcal{Q}(\boldsymbol{S})$	tangent space to the group-sparse matrix variety at point $\boldsymbol{S}$
$\mathrm{gdeg}_{\max}(\boldsymbol{S})$	maximum group degree of group-structured matrix $\boldsymbol{S}$
$\mathrm{gsupp}(\boldsymbol{S})$	group support of group-structured matrix $\boldsymbol{S}$ (indices of non-zero groups)
$\mathrm{gsign}(\boldsymbol{S})$	group-sign function
$\mathcal{L}(r)$	low-rank matrix variety (matrices with at most rank r)
$\mathcal{T}(\boldsymbol{L})$	tangent space to the low-rank matrix variety at point $\boldsymbol{L}$
$\mathrm{coh}(\boldsymbol{L})$	coherence of (the row-/column spaces of) a matrix $\boldsymbol{L}$
ξ, μ	norm-compatibility constants of the $\ell_{\infty,2}$- and spectral norms
$\boldsymbol{x}$	vector of discrete variables
$\boldsymbol{y}$	vector of continuous variables
$\overline{\boldsymbol{x}}$	indicator-coded vector $\boldsymbol{x}$
ℓ	negative log-likelihood (regular log-likelihood in Sections 3.1.2 and C.1)
ℓ_p	negative pseudo log-likelihood
a	log-partition/normalization function
$\hat{\boldsymbol{\Sigma}}$	empirical second-moment matrix
$H(p)$	discrete (Shannon) entropy of distribution p
$\mathbb{E}$	expectation operator
$P_{\mathcal{Q}}$	orthogonal projection on subspace $\mathcal{Q}$
$\rho(\mathcal{T}, \mathcal{T}')$	twisting between subspaces

Appendix B

Additional Material for Chapter 2

B.1 Tangent Spaces and Projections

Low-rank tangent spaces play a fundamental role throughout. Therefore, we characterize the tangent spaces at smooth points of the low-rank matrix variety.

Lemma B.1. *Suppose $\boldsymbol{L} \in \mathcal{L}(r)$ is a rank-r matrix. Then, the tangent space to $\mathcal{L}(r)$ at $\boldsymbol{L}$ is given by*

$$\mathcal{T}(\boldsymbol{L}) = \left\{\boldsymbol{U}\boldsymbol{X}^\mathsf{T} + \boldsymbol{Y}\boldsymbol{V}^\mathsf{T} :\ \boldsymbol{X} \in \mathbb{R}^{n\times r}, \boldsymbol{Y} \in \mathbb{R}^{m\times r}\right\} \subset \mathbb{R}^{m\times n},$$

where $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{E}\boldsymbol{V}^\mathsf{T}$ is the (restricted) singular value decomposition of $\boldsymbol{L}$, that is, $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{n\times r}$ have orthonormal columns and $\boldsymbol{E} \in \mathbb{R}^{r\times r}$ is a diagonal matrix, where the diagonal elements are the non-zero singular values of $\boldsymbol{L}$.

Proof. The tangent space at $\boldsymbol{L}$ is given by the span of all tangent vectors at 0 to smooth curves $\gamma : (-1, 1) \to \mathcal{L}(r)$ initialized at $\boldsymbol{L}$, that is, with $\gamma(0) = \boldsymbol{L}$. Because $\boldsymbol{L}$ is of rank r, it is a smooth point of $\mathcal{L}(r)$ and we can write $\gamma(t) = \boldsymbol{U}(t)\operatorname{sign}(\boldsymbol{E})\boldsymbol{V}(t)^\mathsf{T}$, where for each $t \in (-1, 1)$ the matrices $\boldsymbol{U}(t) \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V}(t) \in \mathbb{R}^{m\times r}$ have rank r, and $\operatorname{sign}(\boldsymbol{E}) \in \mathbb{R}^{r\times r}$ is the diagonal matrix with the signs of the non-zero singular values of $\boldsymbol{L}$ on its diagonal, that is, the diagonal entries are in $\{-1, 1\}$. We can assume the signs of the singular values along the curve to be fixed because we only consider smooth curves. Note that because of $\boldsymbol{L} = \gamma(0) = \boldsymbol{U}(0)\operatorname{sign}(\boldsymbol{E})\boldsymbol{V}(0)^\mathsf{T}$, we can assume w.l.og. that $\boldsymbol{U}(0) = \boldsymbol{U}|\boldsymbol{E}|^{1/2}$ and $\boldsymbol{V}(0) = \boldsymbol{V}|\boldsymbol{E}|^{1/2}$. Hence, applying the chain rule yields

$$\begin{aligned}\gamma'(0) &= \boldsymbol{U}(0)\operatorname{sign}(\boldsymbol{E})\boldsymbol{V}'(0)^\mathsf{T} + \boldsymbol{U}'(0)\operatorname{sign}(\boldsymbol{E})\boldsymbol{V}(0)^\mathsf{T}\\ &= \boldsymbol{U}|\boldsymbol{E}|^{1/2}\operatorname{sign}(\boldsymbol{E})\boldsymbol{V}'(0)^\mathsf{T} + \boldsymbol{U}'(0)\operatorname{sign}(\boldsymbol{E})|\boldsymbol{E}|^{1/2}\boldsymbol{V}^\mathsf{T}.\end{aligned}$$

Setting $\boldsymbol{X} = \boldsymbol{V}'(0)\operatorname{sign}(\boldsymbol{E})|\boldsymbol{E}|^{1/2}$ and $\boldsymbol{Y} = \boldsymbol{U}'(0)\operatorname{sign}(\boldsymbol{E})|\boldsymbol{E}|^{1/2}$, the tangent vector clearly is of the form $\boldsymbol{U}\boldsymbol{X}^\mathsf{T} + \boldsymbol{Y}\boldsymbol{V}^\mathsf{T}$. We still need to show that $\boldsymbol{X}$ and $\boldsymbol{Y}$ can take any values. For that, let $\boldsymbol{X} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{Y} \in \mathbb{R}^{m\times r}$ be arbitrary. Then, consider the specific curve defined by

$$\begin{aligned}\boldsymbol{U}(t) &= \boldsymbol{U}|\boldsymbol{E}|^{1/2} + t\boldsymbol{Y}|\boldsymbol{E}|^{-1/2}\operatorname{sign}(\boldsymbol{E})\\ \boldsymbol{V}(t) &= \boldsymbol{V}|\boldsymbol{E}|^{1/2} + t\boldsymbol{X}|\boldsymbol{E}|^{-1/2}\operatorname{sign}(\boldsymbol{E}).\end{aligned}$$

For sufficiently small t, both $\boldsymbol{U}(t)$ and $\boldsymbol{V}(t)$ have rank r since $\boldsymbol{U}|\boldsymbol{E}|^{1/2}$ and $\boldsymbol{V}|\boldsymbol{E}|^{1/2}$ have rank r and the curve is smooth. Hence, the specific definitions of $\boldsymbol{U}(t)$ and $\boldsymbol{V}(t)$ yield a valid curve. Its derivative at zero computes as

$$\begin{aligned}\gamma'(0) &= \boldsymbol{U}|\boldsymbol{E}|^{1/2}\operatorname{sign}(\boldsymbol{E})\boldsymbol{V}'(0)^\mathsf{T} + \boldsymbol{U}'(0)\operatorname{sign}(\boldsymbol{E})|\boldsymbol{E}|^{1/2}\boldsymbol{V}^\mathsf{T}\\ &= \boldsymbol{U}|\boldsymbol{E}|^{1/2}\operatorname{sign}(\boldsymbol{E})\left(\boldsymbol{X}|\boldsymbol{E}|^{-1/2}\operatorname{sign}(\boldsymbol{E})\right)^\mathsf{T} + \left(\boldsymbol{Y}|\boldsymbol{E}|^{-1/2}\operatorname{sign}(\boldsymbol{E})\right)\operatorname{sign}(\boldsymbol{E})|\boldsymbol{E}|^{1/2}\boldsymbol{V}^\mathsf{T}\\ &= \boldsymbol{U}\boldsymbol{X}^\mathsf{T} + \boldsymbol{Y}\boldsymbol{V}^\mathsf{T}.\end{aligned}$$

This completes the proof. ■

It should be noted that in the case, where the embedding space for the low-rank matrices is $\operatorname{Sym}(w)$ instead of $\mathbb{R}^{m\times n}$, then the tangent space has the following simpler form:

$$\mathcal{T}(\boldsymbol{L}) = \left\{\boldsymbol{U}\boldsymbol{X}^\mathsf{T} + \boldsymbol{X}\boldsymbol{U}^\mathsf{T} :\ \boldsymbol{X} \in \mathbb{R}^{w\times r}\right\} \subset \operatorname{Sym}(w),$$

where now $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{E}\boldsymbol{U}^\mathsf{T}$ is the (restricted) eigenvalue decomposition of $\boldsymbol{L}$.

One consequence of the form of the tangent spaces is the following lemma that concerns the norms of projections on certain tangent spaces and their orthogonal complements.

Lemma B.2. *Let $\mathcal{Q}(\boldsymbol{S})$ be the tangent space at $\boldsymbol{S} \in \mathcal{S}(|\operatorname{gsupp}(\boldsymbol{S})|)$. Then, for any $\boldsymbol{M} \in \mathbb{R}^{m\times n}$, it holds that*

$$\|P_{\mathcal{Q}(\boldsymbol{S})}\boldsymbol{M}\|_{\infty,2} \le \|\boldsymbol{M}\|_{\infty,2} \quad \textit{and} \quad \|P_{\mathcal{Q}(\boldsymbol{S})^\perp}\boldsymbol{M}\|_{\infty,2} \le \|\boldsymbol{M}\|_{\infty,2}.$$

Next, let $\mathcal{T}(\boldsymbol{L})$ be the tangent space at $\boldsymbol{L} \in \mathcal{L}(\operatorname{rank}(\boldsymbol{L}))$. Then, for $\boldsymbol{N} \in \mathbb{R}^{m\times n}$, it holds that

$$\|P_{\mathcal{T}(\boldsymbol{L})}\boldsymbol{N}\| \le 2\|\boldsymbol{N}\| \quad \textit{and} \quad \|P_{\mathcal{T}(\boldsymbol{L})^\perp}\boldsymbol{N}\| \le \|\boldsymbol{N}\|.$$

Proof. We only show the claims concerning the projections on $\mathcal{T}(\boldsymbol{L})$ and $\mathcal{T}^\perp(\boldsymbol{L})$. The other claims are easy. Recall that by Lemma B.1 we have for smooth $\boldsymbol{L} \in \mathcal{L}(r)$ that

$$\mathcal{T}(\boldsymbol{L}) = \left\{\boldsymbol{U}\boldsymbol{X}^\mathsf{T} + \boldsymbol{Y}\boldsymbol{V}^\mathsf{T} :\ \boldsymbol{X} \in \mathbb{R}^{n\times r},\, \boldsymbol{Y} \in \mathbb{R}^{m\times r}\right\},$$

where $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{E}\boldsymbol{V}^\mathsf{T}$ is the (restricted) singular decomposition of $\boldsymbol{L}$. Then, we have more explicitly that

$$P_{\mathcal{T}(\boldsymbol{L})}\boldsymbol{N} = \boldsymbol{P}_U\boldsymbol{N} + \boldsymbol{N}\boldsymbol{P}_V - \boldsymbol{P}_U\boldsymbol{N}\boldsymbol{P}_V = \boldsymbol{P}_U\boldsymbol{N} + (\boldsymbol{I}_m - \boldsymbol{P}_U)\boldsymbol{N}\boldsymbol{P}_V,$$

where $\boldsymbol{I}_m$ is the $(m \times m)$ identity matrix, and $\boldsymbol{P}_U = \boldsymbol{U}\boldsymbol{U}^\mathsf{T}$ and $\boldsymbol{P}_V = \boldsymbol{V}\boldsymbol{V}^\mathsf{T}$ project onto the column spaces of $\boldsymbol{U}$ and $\boldsymbol{V}$, respectively. Note that $P_{\mathcal{T}(\boldsymbol{L})}\boldsymbol{N} \in \mathcal{T}(\boldsymbol{L})$ since

$$\begin{aligned} P_{\mathcal{T}(\boldsymbol{L})}\boldsymbol{N} &= \boldsymbol{P}_U\boldsymbol{N} + (\boldsymbol{I}_m - \boldsymbol{P}_U)\boldsymbol{N}\boldsymbol{P}_V \\ &= \boldsymbol{U}\left(\boldsymbol{U}^\mathsf{T}\boldsymbol{N}\right) + \left((\boldsymbol{I}_m - \boldsymbol{P}_U)\boldsymbol{N}\boldsymbol{V}\right)\boldsymbol{V}^\mathsf{T} = \boldsymbol{U}\boldsymbol{X}^\mathsf{T} + \boldsymbol{Y}\boldsymbol{V}^\mathsf{T} \end{aligned}$$

with $\boldsymbol{X} = \boldsymbol{N}^\mathsf{T}\boldsymbol{U} \in \mathbb{R}^{n\times r}$ and $\boldsymbol{Y} = (\boldsymbol{I}_m - \boldsymbol{P}_U)\boldsymbol{N}\boldsymbol{V} \in \mathbb{R}^{m\times r}$. Moreover, $\boldsymbol{N} - P_{\mathcal{T}(\boldsymbol{L})}\boldsymbol{N}$ is orthogonal to $\mathcal{T}(\boldsymbol{L})$ since

$$\boldsymbol{N} - P_{\mathcal{T}(\boldsymbol{L})}\boldsymbol{N} = \boldsymbol{N} - \boldsymbol{P}_U\boldsymbol{N} - \boldsymbol{N}\boldsymbol{P}_V + \boldsymbol{P}_U\boldsymbol{N}\boldsymbol{P}_V = (\boldsymbol{I}_m - \boldsymbol{P}_U)\boldsymbol{N}(\boldsymbol{I}_n - \boldsymbol{P}_V),$$

and since for any $\boldsymbol{U}\boldsymbol{X}^\mathsf{T} + \boldsymbol{Y}\boldsymbol{V}^\mathsf{T} \in \mathcal{T}(\boldsymbol{L})$ we have

$$\begin{aligned} &\left\langle (\boldsymbol{I}_m - \boldsymbol{P}_U)\boldsymbol{N}(\boldsymbol{I}_n - \boldsymbol{P}_V), \boldsymbol{U}\boldsymbol{X}^\mathsf{T} + \boldsymbol{Y}\boldsymbol{V}^\mathsf{T} \right\rangle \\ &\quad = \operatorname{tr}\left((\boldsymbol{I}_n - \boldsymbol{P}_V)\boldsymbol{N}^\mathsf{T}(\boldsymbol{I}_m - \boldsymbol{P}_U)\boldsymbol{U}\boldsymbol{X}^\mathsf{T}\right) + \operatorname{tr}\left((\boldsymbol{I}_n - \boldsymbol{P}_V)\boldsymbol{N}^\mathsf{T}(\boldsymbol{I}_m - \boldsymbol{P}_U)\boldsymbol{Y}\boldsymbol{V}^\mathsf{T}\right) \\ &\quad = \operatorname{tr}\left((\boldsymbol{I}_n - \boldsymbol{P}_V)\boldsymbol{N}^\mathsf{T}(\boldsymbol{I}_m - \boldsymbol{P}_U)\boldsymbol{Y}\boldsymbol{V}^\mathsf{T}\right) \\ &\quad = \operatorname{tr}\left(\boldsymbol{Y}\boldsymbol{V}^\mathsf{T}(\boldsymbol{I}_n - \boldsymbol{P}_V)\boldsymbol{N}^\mathsf{T}(\boldsymbol{I}_m - \boldsymbol{P}_U)\right) \\ &\quad = 0, \end{aligned}$$

where the second equality follow from $(\boldsymbol{I}_m - \boldsymbol{P}_U)\boldsymbol{U}\boldsymbol{X}^\mathsf{T} = (\boldsymbol{I}_m - \boldsymbol{U}\boldsymbol{U}^\mathsf{T})\boldsymbol{U}\boldsymbol{X}^\mathsf{T} = \boldsymbol{U}\boldsymbol{X}^\mathsf{T} - \boldsymbol{U}\boldsymbol{X}^\mathsf{T} = \boldsymbol{0}$, the third equality uses $\operatorname{tr}(\boldsymbol{A}\boldsymbol{B}) = \operatorname{tr}(\boldsymbol{B}\boldsymbol{A})$, and the last equality follows from $\boldsymbol{Y}\boldsymbol{V}^\mathsf{T}(\boldsymbol{I}_n - \boldsymbol{P}_V) = \boldsymbol{Y}\boldsymbol{V}^\mathsf{T}(\boldsymbol{I}_n - \boldsymbol{V}\boldsymbol{V}^\mathsf{T}) = \boldsymbol{Y}\boldsymbol{V}^\mathsf{T} - \boldsymbol{Y}\boldsymbol{V}^\mathsf{T} = \boldsymbol{0}$. Thus, $P_{\mathcal{T}(\boldsymbol{L})}\boldsymbol{N}$ is indeed the orthogonal projection of $\boldsymbol{N}$ onto $\mathcal{T}(\boldsymbol{L})$. Now, by submultiplicativity of the spectral norm

$$\begin{aligned} \|P_{\mathcal{T}(\boldsymbol{L})}\boldsymbol{N}\| &\le \|\boldsymbol{P}_U\boldsymbol{N}\| + \|(\boldsymbol{I}_m - \boldsymbol{P}_U)\boldsymbol{N}\boldsymbol{P}_V\| \\ &\le \|\boldsymbol{P}_U\|\|\boldsymbol{N}\| + \|(\boldsymbol{I}_m - \boldsymbol{P}_U)\|\|\boldsymbol{N}\|\|\boldsymbol{P}_V\| \le 2\|\boldsymbol{N}\| \end{aligned}$$

because $\boldsymbol{P}_U$, $\boldsymbol{P}_V$, and $\boldsymbol{I}_m - \boldsymbol{P}_U$ are projection matrices, that is, their operator norm is bounded by one. Likewise, it holds

$$\begin{aligned} \|P_{\mathcal{T}^\perp(\boldsymbol{L})}\boldsymbol{N}\| = \|\boldsymbol{N} - P_{\mathcal{T}(\boldsymbol{L})}\boldsymbol{N}\| &= \|(\boldsymbol{I}_m - \boldsymbol{P}_U)\boldsymbol{N}(\boldsymbol{I}_n - \boldsymbol{P}_V)\| \\ &\le \|(\boldsymbol{I}_m - \boldsymbol{P}_U)\|\|\boldsymbol{N}\|\|(\boldsymbol{I}_n - \boldsymbol{P}_V)\| \le \|\boldsymbol{N}\|. \end{aligned}$$

This concludes the proof. ■

B.2 Proof of Lemma 2.1: Local Identifiability

Proof of Lemma 2.1. To prove local identifiability, we must find a (small) ball such that for all $\boldsymbol{\Delta} \neq \mathbf{0}$ from this ball it holds that $(\boldsymbol{L} - \boldsymbol{\Delta}, \boldsymbol{S} + \boldsymbol{\Delta}) \notin \mathcal{L}(r) \times \mathcal{S}(s)$. Recall that the points that are close to $\boldsymbol{S} \in \mathcal{S}(s)$ and $\boldsymbol{L} \in \mathcal{L}(r)$ in the varieties can be characterized using tangent spaces. Specifically, it can only hold $\boldsymbol{S} + \boldsymbol{\Delta} \in \mathcal{S}(s)$ for small $\boldsymbol{\Delta} \neq \mathbf{0}$ *if* $\boldsymbol{\Delta} \in \mathcal{Q}(\boldsymbol{S})$. Likewise, it can only hold $\boldsymbol{L} - \boldsymbol{\Delta} \in \mathcal{L}(r)$ for small $\boldsymbol{\Delta} \neq \mathbf{0}$ *if* $\boldsymbol{\Delta} \in \mathcal{T}(\boldsymbol{L}')$ for some tangent space $\mathcal{T}(\boldsymbol{L}')$ to $\mathcal{L}(r)$ at a (smooth) point $\boldsymbol{L}' \in \mathcal{L}(r)$ that is close to $\boldsymbol{L}$. Note again that due to the local curvature of the low-rank matrix variety, we also need to consider nearby tangent spaces. Hence, to prove local identifiability, it is sufficient to show that the tangent spaces $\mathcal{T}(\boldsymbol{L}')$ and $\mathcal{Q}(\boldsymbol{S})$ are transverse for all smooth $\boldsymbol{L}' \in \mathcal{L}(r)$ from some small ball around $\boldsymbol{L}$. Here, by definition, the transversality of $\mathcal{T}(\boldsymbol{L}')$ and $\mathcal{Q}(\boldsymbol{S})$ means that $\mathcal{T}(\boldsymbol{L}') \cap \mathcal{Q}(\boldsymbol{S}) = \{\mathbf{0}\}$, which is equivalent to

$$\min_{M \in \mathcal{Q}(\boldsymbol{S}),\, \|\boldsymbol{M}\|=1} \|\boldsymbol{M} - P_{\mathcal{T}(\boldsymbol{L}')}\boldsymbol{M}\| > 0. \tag{B.1}$$

This is because $\boldsymbol{M} = P_{\mathcal{T}(\boldsymbol{L}')}\boldsymbol{M}$ if and only if $\boldsymbol{M} \in \mathcal{T}(\boldsymbol{L}') \cap \mathcal{Q}(\boldsymbol{S})$. In the following, we want to verify Condition (B.1) for smooth $\boldsymbol{L}' \in \mathcal{L}(r)$ from a small ball around $\boldsymbol{L}$. We start by calculating that for any $\boldsymbol{M} \in \mathcal{Q}(\boldsymbol{S})$ with $\|\boldsymbol{M}\| = 1$ it holds that

$$\begin{aligned}
\|\boldsymbol{M} - P_{\mathcal{T}(\boldsymbol{L}')}\boldsymbol{M}\| &= \|\boldsymbol{M} - P_{\mathcal{T}(\boldsymbol{L})}\boldsymbol{M} + \left[P_{\mathcal{T}(\boldsymbol{L})} - P_{\mathcal{T}(\boldsymbol{L}')}\right]\boldsymbol{M}\| \\
&\geq \|\boldsymbol{M} - P_{\mathcal{T}(\boldsymbol{L})}\boldsymbol{M}\| - \|\left[P_{\mathcal{T}(\boldsymbol{L})} - P_{\mathcal{T}(\boldsymbol{L}')}\right]\boldsymbol{M}\| \\
&\geq \kappa - \rho(\mathcal{T}(\boldsymbol{L}), \mathcal{T}(\boldsymbol{L}')),
\end{aligned}$$

where the first inequality is the triangle inequality, and for the second inequality we defined

$$\kappa = \min_{M \in \mathcal{Q}(\boldsymbol{S}),\, \|\boldsymbol{M}\|=1} \|\boldsymbol{M} - P_{\mathcal{T}(\boldsymbol{L})}\boldsymbol{M}\|$$

and the twisting between subspaces

$$\rho(\mathcal{T}(\boldsymbol{L}), \mathcal{T}(\boldsymbol{L}')) = \max_{\|\boldsymbol{M}\|=1} \left\|\left[P_{\mathcal{T}(\boldsymbol{L})} - P_{\mathcal{T}(\boldsymbol{L}')}\right]\boldsymbol{M}\right\|.$$

Here, the assumed transversality of the tangent spaces $\mathcal{T}(\boldsymbol{L})$ and $\mathcal{Q}(\boldsymbol{S})$ implies that $\kappa > 0$. Hence, a sufficient condition for the transversality of $\mathcal{Q}(\boldsymbol{S})$ and $\mathcal{T}(\boldsymbol{L}')$ is that $\rho(\mathcal{T}(\boldsymbol{L}), \mathcal{T}(\boldsymbol{L}')) < \kappa$ since then Condition (B.1) is satisfied. Thus, our goal is to show that $\rho(\mathcal{T}(\boldsymbol{L}), \mathcal{T}(\boldsymbol{L}')) < \kappa$ holds whenever $\boldsymbol{L}'$ is sufficiently close to $\boldsymbol{L}$. The proof is technical, but the main idea is to show that the map from smooth $\boldsymbol{L}' \in \mathcal{L}(r)$ to $\rho(\mathcal{T}(\boldsymbol{L}), \mathcal{T}(\boldsymbol{L}'))$ is continuous and since it maps $\boldsymbol{L}$ onto zero, there exists a small ball around $\boldsymbol{L}$ for which $\rho(\mathcal{T}(\boldsymbol{L}), \mathcal{T}(\boldsymbol{L}')) < \kappa$.

We now dive into the technical details. For that, we consider the function f that maps $(\boldsymbol{L}', \boldsymbol{M})$ with domain restricted to $\boldsymbol{L}' \in \mathcal{L}(r)$, $\|\boldsymbol{L} - \boldsymbol{L}'\| \leq 1$, and $\|\boldsymbol{M}\| = 1$

onto $\mathbb{R}$ as follows

$$f(\boldsymbol{L}', \boldsymbol{M}) = \|P_{\mathcal{T}(\boldsymbol{L})}\boldsymbol{M} - (P_{U(\boldsymbol{L}')}\boldsymbol{M} + \boldsymbol{M}P_{V(\boldsymbol{L}')} - P_{U(\boldsymbol{L}')}\boldsymbol{M}P_{V(\boldsymbol{L}')})\|,$$

where $P_{U(\boldsymbol{L}')}$ is the projection matrix that projects onto the column space $U(\boldsymbol{L}')$ of $\boldsymbol{L}'$, and $P_{V(\boldsymbol{L}')}$ is the projection matrix that projects onto the row space $V(\boldsymbol{L}')$ of $\boldsymbol{L}'$. Note that for a rank-r matrix $\boldsymbol{L}'$, which is a smooth point in $\mathcal{L}(r)$, it holds that

$$P_{\mathcal{T}(\boldsymbol{L}')}\boldsymbol{M} = P_{U(\boldsymbol{L}')}\boldsymbol{M} + \boldsymbol{M}P_{V(\boldsymbol{L}')} - P_{U(\boldsymbol{L}')}\boldsymbol{M}P_{V(\boldsymbol{L}')},$$

see the proof of Lemma B.2. Consequently, we have $f(\boldsymbol{L}', \boldsymbol{M}) = \|[P_{\mathcal{T}(\boldsymbol{L})} - P_{\mathcal{T}(\boldsymbol{L}')}]\boldsymbol{M}\|$ for smooth $\boldsymbol{L}'$ and in particular $f(\boldsymbol{L}, \boldsymbol{M}) = 0$ for all $\boldsymbol{M}$. We now argue that f is continuous as a composition of continuous functions: First, $\boldsymbol{L}'$ maps continuously onto the projection matrices $P_{U(\boldsymbol{L}')}, P_{V(\boldsymbol{L}')}$ because small changes to $\boldsymbol{L}'$ only cause small changes to the row and column spaces of $\boldsymbol{L}'$ and hence to the corresponding projections. Second, the remaining composite functions in the definition of f above are additions, norm functions, or matrix products of $P_{U(\boldsymbol{L}')}$, $P_{V(\boldsymbol{L}')}$, and $\boldsymbol{M}$. All these operations are continuous, thus overall f is continuous.

Because f is continuous on a compact domain, it is also uniformly continuous. Hence, there exists $\delta > 0$ (w.l.o.g. $\delta \le 1$) such that for all $\boldsymbol{L}_1', \boldsymbol{L}_2'$ with $\|\boldsymbol{L}_1' - \boldsymbol{L}_2'\| < \delta$ and for all $\boldsymbol{M}_1, \boldsymbol{M}_2$ with $\|\boldsymbol{M}_1 - \boldsymbol{M}_2\| < \delta$ it holds that $|f(\boldsymbol{L}_1', \boldsymbol{M}_1) - f(\boldsymbol{L}_2', \boldsymbol{M}_2)| < \kappa/2$. Consequently, it holds for $\boldsymbol{L}'$ with $\|\boldsymbol{L} - \boldsymbol{L}'\| < \delta$ independently of $\boldsymbol{M}$ that

$$f(\boldsymbol{L}', \boldsymbol{M}) < f(\boldsymbol{L}, \boldsymbol{M}) + \frac{\kappa}{2} = \frac{\kappa}{2}.$$

We can take the supremum over $\boldsymbol{M}$ with $\|\boldsymbol{M}\| = 1$ on the left-hand side of this equation. If we only consider smooth $\boldsymbol{L}'$, this implies that

$$\rho(\mathcal{T}(\boldsymbol{L}), \mathcal{T}(\boldsymbol{L}')) = \sup_{\boldsymbol{M} :\, \|\boldsymbol{M}\|=1} \|[P_{\mathcal{T}(\boldsymbol{L})} - P_{\mathcal{T}(\boldsymbol{L}')}]\boldsymbol{M}\| = \sup_{\boldsymbol{M} :\, \|\boldsymbol{M}\|=1} f(\boldsymbol{L}', \boldsymbol{M}) \le \frac{\kappa}{2} < \kappa.$$

This completes the proof because we have shown that for all smooth $\boldsymbol{L}'$ from the spectral-norm ball with radius δ around $\boldsymbol{L}$ the tangent spaces $\mathcal{T}(\boldsymbol{L}')$ and $\mathcal{Q}(\boldsymbol{S})$ are transverse. Particularly, there do not exist small non-zero $\boldsymbol{\Delta} \in \mathcal{T}(\boldsymbol{L}') \cap \mathcal{Q}(\boldsymbol{S})$ for any $\boldsymbol{L}'$ from that ball, hence locally around $(\boldsymbol{L}, \boldsymbol{S})$ there are no alternative decompositions $(\boldsymbol{L} - \boldsymbol{\Delta}, \boldsymbol{S} + \boldsymbol{\Delta}) \in \mathcal{L}(r) \times \mathcal{S}(s)$. This establishes local identifiability of $(\boldsymbol{L}, \boldsymbol{S})$ in $\mathcal{L}(r) \times \mathcal{S}(s)$. ■

B.3 Proof of Lemma 2.3

Proof of Lemma 2.3. We prove the bound for $\mu(\mathcal{Q}(\boldsymbol{S}))$ first. Remember the definition

$$\mu(\mathcal{Q}(\boldsymbol{S})) = \max_{M \in \mathcal{Q}(\boldsymbol{S}),\, \|\boldsymbol{M}\|_{\infty,2}=1} \|\boldsymbol{M}\|,$$

and recall that the *group-sign* function gsign maps a matrix $\boldsymbol{A} \in \mathbb{R}^{m\times n}$ onto the matrix $\operatorname{gsign}(\boldsymbol{A}) \in \mathbb{R}^{m\times n}$ with

$$\operatorname{gsign}(\boldsymbol{A})_{ij} = \begin{cases} \boldsymbol{a}_{ij}/\|\boldsymbol{a}_{ij}\|_2, & \boldsymbol{a}_{ij} \neq \boldsymbol{0} \\ \boldsymbol{0}, & \text{else} \end{cases}, \qquad i \in 1, \dots, d \quad \text{and} \quad j = 1, \dots, n.$$

Now, consider $\boldsymbol{M}$ with $\|\boldsymbol{M}\|_{\infty,2} = 1$. Then, the matrix $|\operatorname{gsign}(\boldsymbol{M})|$ has normalized groups and non-negative entries. Moreover, it satisfies the element-wise inequality $|\boldsymbol{M}| \leq |\operatorname{gsign}(\boldsymbol{M})|$. As a consequence of the Perron-Frobenius theorem [Horn and Johnson, 2012] it follows that $\|\boldsymbol{M}\| \leq \||\operatorname{gsign}(\boldsymbol{M})|\|$. This allows us to conclude that to compute $\mu(\mathcal{Q}(\boldsymbol{S}))$ we only need to consider non-negative matrices $\boldsymbol{0} \leq \boldsymbol{M} \in \mathcal{Q}(\boldsymbol{S})$ that are *normalized* in the sense that its non-zero groups precisely have norm 1, that is, $\boldsymbol{M} = \operatorname{gsign}(\boldsymbol{M})$. Now, we bound the spectral norm of a matrix $\boldsymbol{M}$, see [Schur, 1911], as follows

$$\|\boldsymbol{M}\|^2 \leq \|\boldsymbol{M}\|_1 \|\boldsymbol{M}\|_\infty,$$

where $\|\boldsymbol{M}\|_1$ is the maximum ℓ_1-norm of a column of $\boldsymbol{M}$ and $\|\boldsymbol{M}\|_\infty$ is the maximum ℓ_1-norm of a row of $\boldsymbol{M}$. We bound $\|\boldsymbol{M}\|_1$. W.l.o.g. let $\boldsymbol{c}$ be a column of $\boldsymbol{M}$ with $\|\boldsymbol{M}\|_1 = \|\boldsymbol{c}\|_1$. Then, it holds

$$\|\boldsymbol{M}\|_1 = \|\boldsymbol{c}\|_1 \leq \sqrt{\eta}\|\boldsymbol{c}\|_{1,2} \leq \sqrt{\eta}\operatorname{gdeg}_{\max}(\boldsymbol{M}) \leq \sqrt{\eta}\operatorname{gdeg}_{\max}(\boldsymbol{S}),$$

where the first inequality follows since $\eta = \max_{i\in[d]} m_i$ is the maximum number of elements that belong to a group from the column and hence that $\sqrt{\eta}$ is a norm compatibility constant for the column (vector) ℓ_1-norm and the column (vector) $\ell_{1,2}$-norm. The second inequality follows because the vector $\ell_{1,2}$-norm of c is equal to the number of non-zero groups of c. This number is bounded by $\operatorname{gdeg}_{\max}(\boldsymbol{M})$. Finally, the last inequality follows from $\operatorname{gsupp}(\boldsymbol{M}) \subseteq \operatorname{gsupp}(\boldsymbol{S})$ as $\boldsymbol{M} \in \mathcal{Q}(\boldsymbol{S})$.

Similar reasoning for rows instead of columns leads us to conclude that $\|\boldsymbol{M}\|_\infty \leq \operatorname{gdeg}_{\max}(\boldsymbol{S})$ since this time comparison of the $\ell_{1,2}$- and ℓ_1 row (vector) norms is not necessary because the intersections of the groups of $\boldsymbol{M}$ with a row respectively contain at most one element. Therefore, we get the upper bound

$$\|\boldsymbol{M}\| \leq \sqrt{\|\boldsymbol{M}\|_1\|\boldsymbol{M}\|_\infty} \leq \eta^{1/4}\operatorname{gdeg}_{\max}(\boldsymbol{S}).$$

This establishes the claim $\mu(Q(\boldsymbol{S})) \leq \eta^{1/4}\operatorname{gdeg}_{\max}(\boldsymbol{S})$.

Proof of (b). The claim about $\xi(\mathcal{T}(\boldsymbol{L}))$ follows from

$$\xi(\mathcal{T}) = \max_{\boldsymbol{M}\in\mathcal{T}(\boldsymbol{L}),\, \|\boldsymbol{M}\|=1} \|\boldsymbol{M}\|_{\infty,2} \leq \sqrt{\eta} \max_{\boldsymbol{M}\in\mathcal{T}(\boldsymbol{L}),\, \|\boldsymbol{M}\|=1} \|\operatorname{vec}(\boldsymbol{M})\|_\infty \leq 2\sqrt{\eta}\operatorname{coh}(\boldsymbol{L}),$$

where the first inequality follows from the general comparison of the (vector) ℓ_∞-norm and the $\ell_{\infty,2}$-norm which holds because η is the maximum number of elements of a group. Finally, the last inequality is a consequence of [Chandrasekaran et al., 2011, Proposition 4]. This finishes the proof. ■

B.4 Proof of Exact Recovery

In the next sections, we prove Theorem 2.6 by studying the optimality conditions of Problem (2.2).

B.4.1 Optimality conditions

Before we prove Proposition 2.5, we show a simple claim.

Lemma B.3 (Hoelder-like inequalities for dual norm pairs)**.** *Let $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$. Then, it holds for any pair of dual norms $(\|\cdot\|, \|\cdot\|^*)$ that*

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| \le \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|^*.$$

Proof. This follows quite straightforward from the definition of the dual norm:

$$|\langle \boldsymbol{x}, \boldsymbol{y} \rangle| = \|\boldsymbol{x}\| \cdot |\langle \boldsymbol{x}/\|\boldsymbol{x}\|, \boldsymbol{y} \rangle| \le \|\boldsymbol{x}\| \cdot \sup\{\langle \boldsymbol{y}, \boldsymbol{z} \rangle \,:\, \|\boldsymbol{z}\| \le 1\} = \|\boldsymbol{x}\| \cdot \|\boldsymbol{y}\|^*.$$

This concludes the proof. ■

Proof of Proposition 2.5. First, it follows that $(\boldsymbol{L}^\star, \boldsymbol{S}^\star)$ is *an* optimum since by the second condition from the assumption there exists a dual $\boldsymbol{Z}$ that satisfies both optimality conditions. Now, for some matrix $\boldsymbol{\Delta}$, let $(\boldsymbol{L}^\star - \boldsymbol{\Delta}, \boldsymbol{S}^\star + \boldsymbol{\Delta})$ be another minimizer of Problem (2.2). The minimizer must have this form in order to be feasible. Our goal is to show that the components of $\boldsymbol{\Delta}$ in the normal spaces $\mathcal{Q}^\perp$ and $\mathcal{T}^\perp$ vanish, respectively (recall that we write $\mathcal{Q} = \mathcal{Q}(\boldsymbol{S}^\star)$ and $\mathcal{T} = \mathcal{T}(\boldsymbol{L}^\star)$). We begin by using the subgradient property:

$$\begin{aligned} 0 &= \|\boldsymbol{L}^\star - \boldsymbol{\Delta}\|_* + \gamma \|\boldsymbol{S}^\star + \boldsymbol{\Delta}\|_{1,2} - \|\boldsymbol{L}^\star\|_* - \gamma \|\boldsymbol{S}^\star\|_{1,2} \\ &\ge \langle \boldsymbol{Z}_{1,2}, \boldsymbol{\Delta} \rangle - \langle \boldsymbol{Z}_*, \boldsymbol{\Delta} \rangle \\ &= \langle P_{\mathcal{Q}^\perp}(\boldsymbol{Z}_{1,2}), \boldsymbol{\Delta} \rangle - \langle P_{\mathcal{T}^\perp}(\boldsymbol{Z}_*), \boldsymbol{\Delta} \rangle + \langle P_{\mathcal{Q}}(\boldsymbol{Z}_{1,2}), \boldsymbol{\Delta} \rangle - \langle P_{\mathcal{T}}(\boldsymbol{Z}_*), \boldsymbol{\Delta} \rangle , \end{aligned}$$

where $\boldsymbol{Z}_{1,2} \in \gamma\, \partial \|\boldsymbol{S}^\star\|_{1,2}$ and $\boldsymbol{Z}_* \in \partial \|\boldsymbol{L}^\star\|_*$ are subgradients whose choices we make precise later. The idea is to chose them such that the right hand side of the inequality is maximized. In the last line, we decomposed the terms into their tangential and normal components. Note that the tangential components $\langle P_{\mathcal{Q}}(\boldsymbol{Z}_{1,2}), \boldsymbol{\Delta} \rangle - \langle P_{\mathcal{T}}(\boldsymbol{Z}_*), \boldsymbol{\Delta} \rangle$ do not depend on the choice of the subgradients since by the subgradient characterizations the pair $(\boldsymbol{Z}_{1,2}, \boldsymbol{Z}_*)$ must satisfy

$$P_{\mathcal{Q}}(\boldsymbol{Z}_{1,2}) = \gamma \operatorname{gsign}(\boldsymbol{S}^\star) \quad \text{and} \quad P_{\mathcal{T}}(\boldsymbol{Z}_*) = \boldsymbol{U}\boldsymbol{V}^\mathsf{T},$$

where $\boldsymbol{L}^\star = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\mathsf{T}$ is a singular value decomposition of $\boldsymbol{L}^\star$. Hence, this constant part can be bounded by

$$\begin{aligned}\langle P_\mathcal{Q}(\boldsymbol{Z}_{1,2}), \boldsymbol{\Delta}\rangle - \langle P_\mathcal{T}(\boldsymbol{Z}_*), \boldsymbol{\Delta}\rangle &= \langle \boldsymbol{Z} - P_{\mathcal{Q}^\perp}(\boldsymbol{Z}), \boldsymbol{\Delta}\rangle - \langle \boldsymbol{Z} - P_{\mathcal{T}^\perp}(\boldsymbol{Z}), \boldsymbol{\Delta}\rangle \\ &= - \langle P_{\mathcal{Q}^\perp}(\boldsymbol{Z}), \boldsymbol{\Delta}\rangle + \langle P_{\mathcal{T}^\perp}(\boldsymbol{Z}), \boldsymbol{\Delta}\rangle \\ &= - \langle P_{\mathcal{Q}^\perp}(\boldsymbol{Z}), P_{\mathcal{Q}^\perp}(\boldsymbol{\Delta})\rangle + \langle P_{\mathcal{T}^\perp}(\boldsymbol{Z}), P_{\mathcal{T}^\perp}(\boldsymbol{\Delta})\rangle \\ &\geq - |\langle P_{\mathcal{Q}^\perp}(\boldsymbol{Z}), P_{\mathcal{Q}^\perp}(\boldsymbol{\Delta})\rangle| - |\langle P_{\mathcal{T}^\perp}(\boldsymbol{Z}), P_{\mathcal{T}^\perp}(\boldsymbol{\Delta})\rangle| \\ &\geq - \|P_{\mathcal{Q}^\perp}(\boldsymbol{Z})\|_{\infty,2} \|P_{\mathcal{Q}^\perp}(\boldsymbol{\Delta})\|_{1,2} - \|P_{\mathcal{T}^\perp}(\boldsymbol{Z})\| \, \|P_{\mathcal{T}^\perp}(\boldsymbol{\Delta})\|_*,\end{aligned}$$

where the first equality uses that $\boldsymbol{Z}$ satisfies the subgradient conditions as well, and the final inequality applies the generalized Hoelder inequality from Lemma B.3 (respectively for the $\|\cdot\|_{1,2}$, $\|\cdot\|_{\infty,2}$ and the $\|\cdot\|$, $\|\cdot\|_*$ dual norm pairs).

Next, we calculate $\langle P_{\mathcal{Q}^\perp}(\boldsymbol{Z}_{1,2}), \boldsymbol{\Delta}\rangle - \langle P_{\mathcal{T}^\perp}(\boldsymbol{Z}_*), \boldsymbol{\Delta}\rangle$ after choosing the normal components of $\boldsymbol{Z}_{1,2}$ and $\boldsymbol{Z}_*$ in $\mathcal{Q}^\perp$ and $\mathcal{T}^\perp$, respectively. First, we select $P_{\mathcal{Q}^\perp}(\boldsymbol{Z}_{1,2}) = \gamma \operatorname{gsign}(P_{\mathcal{Q}^\perp}(\boldsymbol{\Delta}))$. This yields a valid subgradient because then $\|P_{\mathcal{Q}^\perp}(\boldsymbol{Z}_{1,2})\|_{\infty,2} = \gamma$. Moreover, it holds that

$$\begin{aligned}\langle P_{\mathcal{Q}^\perp}(\boldsymbol{Z}_{1,2}), \boldsymbol{\Delta}\rangle = \langle P_{\mathcal{Q}^\perp}(\boldsymbol{Z}_{1,2}), P_{\mathcal{Q}^\perp}(\boldsymbol{\Delta})\rangle &= \gamma \left\langle \operatorname{gsign}(P_{\mathcal{Q}^\perp}(\boldsymbol{\Delta})), P_{\mathcal{Q}^\perp}(\boldsymbol{\Delta})\right\rangle \\ &= \gamma \|P_{\mathcal{Q}^\perp}(\boldsymbol{\Delta})\|_{1,2}.\end{aligned}$$

Second, we select $P_{\mathcal{T}^\perp}(\boldsymbol{Z}_*) = -\tilde{\boldsymbol{U}} \operatorname{sign}(\tilde{\boldsymbol{\Sigma}})\tilde{\boldsymbol{V}}^\mathsf{T}$ based on a singular value decomposition $P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}) = \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^\mathsf{T}$ of $P_{\mathcal{T}^\perp}(\boldsymbol{\Delta})$. This forms a valid subgradient since $\|P_{\mathcal{T}^\perp}(\boldsymbol{Z}_*)\| = 1$. Besides, we have

$$\begin{aligned}- \langle P_{\mathcal{T}^\perp}(\boldsymbol{Z}_*), \boldsymbol{\Delta}\rangle = - \langle P_{\mathcal{T}^\perp}(\boldsymbol{Z}_*), P_{\mathcal{T}^\perp}(\boldsymbol{\Delta})\rangle &= -\langle -\tilde{\boldsymbol{U}} \operatorname{sign}(\tilde{\boldsymbol{\Sigma}})\tilde{\boldsymbol{V}}^\mathsf{T}, \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^\mathsf{T}\rangle \\ &= \operatorname{tr}\left(\left(\tilde{\boldsymbol{U}} \operatorname{sign}(\tilde{\boldsymbol{\Sigma}})\tilde{\boldsymbol{V}}^\mathsf{T}\right)^\mathsf{T} \tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^\mathsf{T}\right) = \operatorname{tr}\left(\tilde{\boldsymbol{V}} \operatorname{sign}(\tilde{\boldsymbol{\Sigma}})\tilde{\boldsymbol{U}}^\mathsf{T}\tilde{\boldsymbol{U}}\tilde{\boldsymbol{\Sigma}}\tilde{\boldsymbol{V}}^\mathsf{T}\right) \\ &= \operatorname{tr}\left(\tilde{\boldsymbol{V}}|\tilde{\boldsymbol{\Sigma}}|\tilde{\boldsymbol{V}}^\mathsf{T}\right) = \operatorname{tr}\left(\tilde{\boldsymbol{V}}^\mathsf{T}\tilde{\boldsymbol{V}}|\tilde{\boldsymbol{\Sigma}}|\right) = \operatorname{tr}\left(|\tilde{\boldsymbol{\Sigma}}|\right) = \|P_{\mathcal{T}^\perp}(\boldsymbol{\Delta})\|_*.\end{aligned}$$

In summary, the specific choices of the subgradients yield

$$\begin{aligned}0 &\geq \langle P_{\mathcal{Q}^\perp}(\boldsymbol{Z}_{1,2}), \boldsymbol{\Delta}\rangle - \langle P_{\mathcal{T}^\perp}(\boldsymbol{Z}_*), \boldsymbol{\Delta}\rangle + \langle P_\mathcal{Q}(\boldsymbol{Z}_{1,2}), \boldsymbol{\Delta}\rangle - \langle P_\mathcal{T}(\boldsymbol{Z}_*), \boldsymbol{\Delta}\rangle \\ &\geq \gamma \|P_{\mathcal{Q}^\perp}(\boldsymbol{\Delta})\|_{1,2} + \|P_{\mathcal{T}^\perp}(\boldsymbol{\Delta})\|_* \\ &\qquad - \|P_{\mathcal{Q}^\perp}(\boldsymbol{Z})\|_{\infty,2} \|P_{\mathcal{Q}^\perp}(\boldsymbol{\Delta})\|_{1,2} - \|P_{\mathcal{T}^\perp}(\boldsymbol{Z})\| \, \|P_{\mathcal{T}^\perp}(\boldsymbol{\Delta})\|_* \\ &= (\gamma - \|P_{\mathcal{Q}^\perp}(\boldsymbol{Z})\|_{\infty,2}) \|P_{\mathcal{Q}^\perp}(\boldsymbol{\Delta})\|_{1,2} + (1 - \|P_{\mathcal{T}^\perp}(\boldsymbol{Z})\|) \|P_{\mathcal{T}^\perp}(\boldsymbol{\Delta})\|_*.\end{aligned}$$

As a consequence of the strictly-dual-feasible condition we have $\|P_{\mathcal{Q}^\perp}(\boldsymbol{Z})\|_{\infty,2} < \gamma$ and $\|P_{\mathcal{T}^\perp}(\boldsymbol{Z})\| < 1$. Thus, if any of $P_{\mathcal{Q}^\perp}(\boldsymbol{\Delta})$ *or* $P_{\mathcal{T}^\perp}(\boldsymbol{\Delta})$ are non-zero, then the right-hand side becomes strictly positive, leading to a contradiction. Therefore, we must have $P_{\mathcal{Q}^\perp}(\boldsymbol{\Delta}) = \boldsymbol{0}$ and $P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}) = \boldsymbol{0}$. This means that $\boldsymbol{\Delta}$ must be contained in both $\mathcal{Q}$ and $\mathcal{T}$. However, since we assumed transversality, we have $\mathcal{Q} \cap \mathcal{T} = \{\boldsymbol{0}\}$. It follows that $\boldsymbol{\Delta} = \boldsymbol{0}$, which implies the uniqueness of the solution to Problem (2.2). ■

B.4.2 Proof of Theorem 2.6: Main result on exact recovery

Here, we prove Theorem 2.6 by showing that for any $\gamma \in (\gamma_{\min}, \gamma_{\max})$ there exists a strictly dual feasible $\boldsymbol{Z}$ as required by Proposition 2.5.

Proof of Theorem 2.6. Let us first check that the range of values for γ given by

$$\left(\frac{\xi(\mathcal{T})}{1-4\xi(\mathcal{T})\mu(\mathcal{Q})}, \frac{1-3\xi(\mathcal{T})\mu(\mathcal{Q})}{\mu(\mathcal{Q})}\right)$$

is non-empty. For that, observe that comparing the borders of the interval leads to the quadratic inequality

$$12\left[\xi(\mathcal{T})\mu(\mathcal{Q})\right]^2 - 8\left[\xi(\mathcal{T})\mu(\mathcal{Q})\right] + 1 > 0$$

in $\xi(\mathcal{T})\mu(\mathcal{Q})$. The roots of the quadratic polynomial are $1/6$ and $1/2$, so clearly under the assumption $\xi(\mathcal{T})\mu(\mathcal{Q}) < 1/6$ the given range is non-empty.

Because of the assumption, we can also apply Lemma 2.4 that yields $\mathcal{T} \cap \mathcal{Q} = \{\boldsymbol{0}\}$. Therefore, there exists a *unique* $\boldsymbol{Z} \in \mathcal{T} \oplus \mathcal{Q}$, where $\oplus$ denotes the direct sum, such that the orthogonal projections of $\boldsymbol{Z}$ onto the tangent spaces $\mathcal{T}$ and $\mathcal{Q}$ are consistent with the subgradient conditions, that is, it holds

$$P_{\mathcal{T}}(\boldsymbol{Z}) = \boldsymbol{U}\boldsymbol{V}^{\mathsf{T}} \quad \text{and} \quad P_{\mathcal{Q}}(\boldsymbol{Z}) = \gamma \operatorname{gsign}(\boldsymbol{S}^{\star}).$$

Remember that $\boldsymbol{L}^{\star} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^{\mathsf{T}}$ is the (restricted) singular value decomposition of $\boldsymbol{L}^{\star}$. The rest of the proof is dedicated to showing that $\boldsymbol{Z}$ also *strictly* satisfies the remaining subgradient conditions that concern the orthogonal projections, that is, we want to show the strict dual-feasible conditions

$$\|P_{\mathcal{T}^{\perp}}(\boldsymbol{Z})\| < 1 \quad \text{and} \quad \|P_{\mathcal{Q}^{\perp}}(\boldsymbol{Z})\|_{\infty,2} < \gamma$$

that are required by Proposition 2.5. For that, let $\boldsymbol{Z} = \boldsymbol{Z}_{\mathcal{T}} + \boldsymbol{Z}_{\mathcal{Q}}$ be the unique splitting of $\boldsymbol{Z}$ into its components $\boldsymbol{Z}_{\mathcal{T}} \in \mathcal{T}$ and $\boldsymbol{Z}_{\mathcal{Q}} \in \mathcal{Q}$, see Figure B.1. We have $\boldsymbol{Z}_{\mathcal{T}} = P_{\mathcal{T}}(\boldsymbol{Z}) - P_{\mathcal{T}}(\boldsymbol{Z}_{\mathcal{Q}}) = \boldsymbol{U}\boldsymbol{V}^{\mathsf{T}} - P_{\mathcal{T}}(\boldsymbol{Z}_{\mathcal{Q}})$ and $\boldsymbol{Z}_{\mathcal{Q}} = P_{\mathcal{Q}}(\boldsymbol{Z}) - P_{\mathcal{Q}}(\boldsymbol{Z}_{\mathcal{T}}) = \gamma \operatorname{gsign}(\boldsymbol{S}^{\star}) - P_{\mathcal{Q}}(\boldsymbol{Z}_{\mathcal{T}})$.

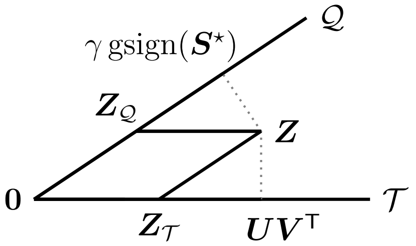
$\mathcal{Q}$
$\gamma \operatorname{gsign}(\boldsymbol{S}^{\star})$
$\boldsymbol{Z}_{\mathcal{Q}}$
$\boldsymbol{Z}$
$\boldsymbol{0}$
$\mathcal{T}$
$\boldsymbol{Z}_{\mathcal{T}}$
$\boldsymbol{U}\boldsymbol{V}^{\mathsf{T}}$

Figure B.1: Decomposition of the dual $\boldsymbol{Z}$ in $\mathcal{Q} \oplus \mathcal{T}$.

Now, we start bounding the orthogonal components. The component of $\boldsymbol{Z}$ in $\mathcal{Q}^\perp$ can be bounded as

$$\begin{aligned}
\|P_{\mathcal{Q}^\perp}(\boldsymbol{Z})\|_{\infty,2} &= \|P_{\mathcal{Q}^\perp}(\boldsymbol{Z}_\mathcal{T})\|_{\infty,2} \leq \|\boldsymbol{Z}_\mathcal{T}\|_{\infty,2} \\
&\leq \xi(\mathcal{T})\|\boldsymbol{Z}_\mathcal{T}\| = \xi(\mathcal{T})\|\boldsymbol{U}\boldsymbol{V}^\mathsf{T} - P_\mathcal{T}(\boldsymbol{Z}_\mathcal{Q})\| \\
&\leq \xi(\mathcal{T})(1 + \|P_\mathcal{T}(\boldsymbol{Z}_\mathcal{Q})\|), \qquad \text{(B.2)}
\end{aligned}$$

where we used the projection Lemma B.2 in the first, the definition of $\xi(\mathcal{T})$ in the second, and the triangle inequality in the last inequality. Similarly, we can bound the component of $\boldsymbol{Z}$ in $\mathcal{T}^\perp$

$$\begin{aligned}
\|P_{\mathcal{T}^\perp}(\boldsymbol{Z})\| &= \|P_{\mathcal{T}^\perp}(\boldsymbol{Z}_\mathcal{Q})\| \leq \|\boldsymbol{Z}_\mathcal{Q}\| \\
&\leq \mu(\mathcal{Q})\, \|\boldsymbol{Z}_\mathcal{Q}\|_{\infty,2} = \mu(\mathcal{Q})\|\gamma \operatorname{gsign}(\boldsymbol{S}^\star) - P_\mathcal{Q}(\boldsymbol{Z}_\mathcal{T})\|_{\infty,2} \\
&\leq \mu(\mathcal{Q})\, \left(\gamma + \|P_\mathcal{Q}(\boldsymbol{Z}_\mathcal{T})\|_{\infty,2}\right), \qquad \text{(B.3)}
\end{aligned}$$

where again Lemma B.2 was used in the first, the definition of $\mu(\mathcal{Q})$ in the second, and finally the triangle inequality in the last inequality. To continue the calculations we bound the norms of $P_\mathcal{T}(\boldsymbol{Z}_\mathcal{Q})$ and $P_\mathcal{Q}(\boldsymbol{Z}_\mathcal{T})$.

$$\begin{aligned}
\|P_\mathcal{T}(\boldsymbol{Z}_\mathcal{Q})\| &\leq 2\|\boldsymbol{Z}_\mathcal{Q}\| \leq 2\mu(\mathcal{Q})\, \left(\gamma + \|P_\mathcal{Q}(\boldsymbol{Z}_\mathcal{T})\|_{\infty,2}\right), \\
\|P_\mathcal{Q}(\boldsymbol{Z}_\mathcal{T})\|_{\infty,2} &\leq \|\boldsymbol{Z}_\mathcal{T}\|_{\infty,2} \leq \xi(\mathcal{T})\left(1 + \|P_\mathcal{T}(\boldsymbol{Z}_\mathcal{Q})\|\right),
\end{aligned}$$

where we used the projection Lemma B.2, bounded $\|\boldsymbol{Z}_\mathcal{T}\|_{\infty,2}$ as in (B.2), and bounded $\|\boldsymbol{Z}_\mathcal{Q}\|$ as in (B.3). Plugging the bounds on $\|P_\mathcal{T}(\boldsymbol{Z}_\mathcal{Q})\|$ and $\|P_\mathcal{Q}(\boldsymbol{Z}_\mathcal{T})\|$ into each other yields

$$\begin{aligned}
\|P_\mathcal{T}(\boldsymbol{Z}_\mathcal{Q})\| &\leq 2\mu(\mathcal{Q})\, \left[\gamma + \xi(\mathcal{T})\left(1 + \|P_\mathcal{T}(\boldsymbol{Z}_\mathcal{Q})\|\right)\right] \qquad \text{and} \\
\|P_\mathcal{Q}(\boldsymbol{Z}_\mathcal{T})\|_{\infty,2} &\leq \xi(\mathcal{T})\left[1 + 2\mu(\mathcal{Q})\, \left(\gamma + \|P_\mathcal{Q}(\boldsymbol{Z}_\mathcal{T})\|_{\infty,2}\right)\right].
\end{aligned}$$

By solving these inequalities for $\|P_\mathcal{T}(\boldsymbol{Z}_\mathcal{Q})\|$ and $\|P_\mathcal{Q}(\boldsymbol{Z}_\mathcal{T})\|_{\infty,2}$, respectively, we obtain

$$\|P_\mathcal{T}(\boldsymbol{Z}_\mathcal{Q})\| \leq \frac{2\gamma\mu(\mathcal{Q}) + 2\xi(\mathcal{T})\mu(\mathcal{Q})}{1 - 2\xi(\mathcal{T})\mu(\mathcal{Q})} \qquad \text{(B.4a)}$$

$$\|P_\mathcal{Q}(\boldsymbol{Z}_\mathcal{T})\|_{\infty,2} \leq \frac{\xi(\mathcal{T}) + 2\gamma\xi(\mathcal{T})\mu(\mathcal{Q})}{1 - 2\xi(\mathcal{T})\mu(\mathcal{Q})} \qquad \text{(B.4b)}$$

Note that the denominators are positive because of

$$\xi(\mathcal{T})\mu(\mathcal{Q}) < 1/6 < 1/2.$$

Bringing (B.2) and (B.4a) together yields

$$\begin{aligned}
\|P_{\mathcal{Q}^\perp}(\boldsymbol{Z})\|_{\infty,2} &\leq \xi(\mathcal{T})(1 + \|P_{\mathcal{T}}(\boldsymbol{Z}_{\mathcal{Q}})\|) \\
&\leq \xi(\mathcal{T}) \left(1 + \frac{2\gamma\mu(\mathcal{Q}) + 2\xi(\mathcal{T})\mu(\mathcal{Q})}{1 - 2\xi(\mathcal{T})\mu(\mathcal{Q})}\right) = \xi(\mathcal{T}) \left(\frac{1 + 2\gamma\mu(\mathcal{Q})}{1 - 2\xi(\mathcal{T})\mu(\mathcal{Q})}\right) \\
&= \left[\xi(\mathcal{T}) \left(\frac{1 + 2\gamma\mu(\mathcal{Q})}{1 - 2\xi(\mathcal{T})\mu(\mathcal{Q})}\right) - \gamma\right] + \gamma \\
&= \left[\frac{\xi(\mathcal{T}) + 2\gamma\xi(\mathcal{T})\mu(\mathcal{Q}) - \gamma + 2\gamma\xi(\mathcal{T})\mu(\mathcal{Q})}{1 - 2\xi(\mathcal{T})\mu(\mathcal{Q})}\right] + \gamma \\
&= \left[\frac{\xi(\mathcal{T}) - \gamma\,(1 - 4\xi(\mathcal{T})\mu(\mathcal{Q}))}{1 - 2\xi(\mathcal{T})\mu(\mathcal{Q})}\right] + \gamma < \gamma,
\end{aligned}$$

where the last inequality holds by the assumption $\gamma > \xi(\mathcal{T})/[1 - 4\xi(\mathcal{T})\mu(\mathcal{Q})]$. Next, by (B.3) and (B.4b) we have

$$\begin{aligned}
\|P_{\mathcal{T}^\perp}(\boldsymbol{Z})\| &\leq \mu(\mathcal{Q})\ (\gamma + \|P_{\mathcal{Q}}(\boldsymbol{Z}_{\mathcal{T}})\|_{\infty,2}) \\
&\leq \mu(\mathcal{Q}) \left(\gamma + \frac{\xi(\mathcal{T}) + 2\gamma\xi(\mathcal{T})\mu(\mathcal{Q})}{1 - 2\xi(\mathcal{T})\mu(\mathcal{Q})}\right) = \mu(\mathcal{Q}) \left(\frac{\gamma + \xi(\mathcal{T})}{1 - 2\xi(\mathcal{T})\mu(\mathcal{Q})}\right) \\
&< \mu(\mathcal{Q}) \left(\frac{[1 - 3\xi(\mathcal{T})\mu(\mathcal{Q})]/\mu(\mathcal{Q}) + \xi(\mathcal{T})}{1 - 2\xi(\mathcal{T})\mu(\mathcal{Q})}\right) \\
&= \frac{1 - 3\mu(\mathcal{Q})\xi(\mathcal{T}) + \mu(\mathcal{Q})\xi(\mathcal{T})}{1 - 2\xi(\mathcal{T})\mu(\mathcal{Q})} = 1,
\end{aligned}$$

where we used the bound $\gamma < [1 - 3\xi(\mathcal{T})\mu(\mathcal{Q})]/\mu(\mathcal{Q})$ from the assumption in the last inequality. This completes the proof. ■

B.5 Weighted Group Norms

In this section, we present some properties of weighted group norms that are necessary to adapt the assumptions and proofs for exact recovery when Problem (2.6) is used instead of Problem (2.2) for learning RPCA models. We do not detail all steps that are necessary to obtain a general result for exact recovery with weighted group norms. However, with the given information, it should be easy to follow and generalize the lines of the proof of Theorem 2.6.

Lemma B.4 (duality of weighted group norms). *The dual norm of the weighted $\ell_{1,2}$-group norm*

$$\|\boldsymbol{S}\|_{1,2}^{\boldsymbol{W}} = \sum_{i,j} w_{ij}\|\boldsymbol{s}_{ij}\|_2, \qquad \boldsymbol{W} = (w_{ij})_{i\in[d],j\in[n]},$$

is given by the weighted $\ell_{\infty,2}$-group norm

$$\|\boldsymbol{Y}\|_{\infty,2}^{\boldsymbol{W}} = \max_{i,j} w_{ij}^{-1}\|\boldsymbol{y}_{ij}\|_2.$$

Remember that we denote the (i, j)-th sub-groups of a group-structured matrix with bold lowercase letters with corresponding subscript.

Proof. The dual norm (of $\boldsymbol{Y} \in \mathbb{R}^{m \times n}$) is defined as $\max_{\|\boldsymbol{S}\|_{1,2}^{\boldsymbol{W}} \leq 1} \langle \boldsymbol{Y}, \boldsymbol{S} \rangle$. First, it holds

$$\begin{aligned}
\max_{\|\boldsymbol{S}\|_{1,2}^{\boldsymbol{W}} \leq 1} \langle \boldsymbol{Y}, \boldsymbol{S} \rangle &= \max_{\|\boldsymbol{S}\|_{1,2}^{\boldsymbol{W}} \leq 1} \sum_{i,j} \langle \boldsymbol{s}_{ij}, \boldsymbol{y}_{ij} \rangle \\
&\leq \max_{\|\boldsymbol{S}\|_{1,2}^{\boldsymbol{W}} \leq 1} \sum_{i,j} \|\boldsymbol{s}_{ij}\|_2 \, \|\boldsymbol{y}_{ij}\|_2 \\
&\leq \max_{\|\boldsymbol{S}\|_{1,2}^{\boldsymbol{W}} \leq 1} \sum_{i,j} \|\boldsymbol{s}_{ij}\|_2 w_{ij} \|\boldsymbol{Y}\|_{\infty,2}^{\boldsymbol{W}} \\
&\leq \|\boldsymbol{Y}\|_{\infty,2}^{\boldsymbol{W}},
\end{aligned}$$

where the second inequality follows from the definition of the weighted $\ell_{1,2}$-group norm. Second, let (i, j) be such that $w_{ij}^{-1} \|\boldsymbol{y}_{ij}\|_2 = \|\boldsymbol{Y}\|_{\infty,2}^{\boldsymbol{W}}$. Let $\boldsymbol{A}$ be zero except in the (i, j)-th group for which $\boldsymbol{A}_{ij} = \boldsymbol{y}_{ij} / (w_{ij} \|\boldsymbol{y}_{ij}\|_2)$. Then, $\boldsymbol{A}$ satisfies $\|\boldsymbol{A}\|_{1,2}^{\boldsymbol{W}} = 1$ and thus

$$\max_{\|\boldsymbol{S}\|_{1,2}^{\boldsymbol{W}} \leq 1} \langle \boldsymbol{Y}, \boldsymbol{S} \rangle \geq \langle \boldsymbol{Y}, \boldsymbol{A} \rangle = w_{ij}^{-1} \|\boldsymbol{y}_{ij}\|_2 = \|\boldsymbol{Y}\|_{\infty,2}^{\boldsymbol{W}}.$$

This completes the proof. ■

Subdifferential. The elements in the subdifferential are given as

$$\partial \|\boldsymbol{S}\|_{1,2}^{\boldsymbol{W}} = \left\{ \boldsymbol{Y} : \langle \boldsymbol{S}, \boldsymbol{Y} \rangle = \|\boldsymbol{S}\|_{1,2}^{\boldsymbol{W}}, \|\boldsymbol{Y}\|_{\infty,2}^{\boldsymbol{W}} \leq 1 \right\}.$$

More precisely, it can be shown that it holds $\boldsymbol{Z} \in \partial \|\boldsymbol{S}\|_{1,2}^{\boldsymbol{W}}$ if and only if

$$P_{\mathcal{Q}(\boldsymbol{S})}(\boldsymbol{Z}) = \boldsymbol{W} \circ \operatorname{gsign}(\boldsymbol{S}) \quad \text{and} \quad \|P_{\mathcal{Q}(\boldsymbol{S})^{\perp}}(\boldsymbol{Z})\|_{\infty,2}^{\boldsymbol{W}} \leq 1,$$

where $\circ$ multiplies w_{ij} to each element of the (i, j)-th group of the operator's right-hand side argument (for all i, j).

Proximal operator. The proximal operator in the ADMM algorithm from Section 2.3 is now solved by the group soft-shrinkage operation that acts on the (i, j)-th group as

$$[\operatorname{gShrink}(\boldsymbol{Z}, \kappa, \boldsymbol{W})]_{ij} = \boldsymbol{z}_{ij} \cdot \max\left\{ 1 - \frac{w_{ij}\kappa}{\|\boldsymbol{z}_{ij}\|_2}, 0 \right\}.$$

Appendix C

Additional Material for Chapter 3

C.1 Entropy and Likelihood Duality

C.1.1 Duality for discrete distributions

In this section, we show that the relaxed maximum-entropy problem

$$\max_{p \in \mathcal{P}} H(p) \quad \text{s.t.} \quad \|\mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}}\|_{\infty,2} \leq c \ \text{ and } \ \|\mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}}\| \leq \lambda, \tag{C.1}$$

from Section 3.1.2 is dual to the following regularized log-likelihood maximization problem

$$\max_{S,\, L \,\in\, \mathrm{Sym}(m)} \ \ell(\boldsymbol{S} + \boldsymbol{L}) - c\|\boldsymbol{S}\|_{1,2} - \lambda\|\boldsymbol{L}\|_{*}.$$

Here, the expectation in the maximum-entropy problem is taken w.r.t. the probability distribution p on $\mathcal{X}$.

Proof. The proof is done using standard duality theory. For that, instead of Problem (C.1) we consider the equivalent augmented problem

$$\begin{aligned}
\min_{\boldsymbol{T} \geq \mathbf{0},\, p \geq \mathbf{0}} \quad -H(p) \quad \text{s.t.} \quad & \|\boldsymbol{T}_{ij}\|_2^2 \leq c^2, && (\gamma_{ij}) \quad \text{for } i, j \in [d],\\
& \hat{\boldsymbol{\Sigma}} - \mathbb{E}[\Sigma] \leq \boldsymbol{T} && (\boldsymbol{S}^+),\\
& \mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}} \leq \boldsymbol{T} && (\boldsymbol{S}^-),\\
& \hat{\boldsymbol{\Sigma}} - \mathbb{E}[\Sigma] \preceq \lambda \boldsymbol{I} && (\boldsymbol{L}_1),\\
& \mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}} \preceq \lambda \boldsymbol{I} && (\boldsymbol{L}_2),\\
& \textstyle\sum_{\boldsymbol{x} \in \mathcal{X}} p(\boldsymbol{x}) = 1 && (\theta_0),
\end{aligned} \tag{C.2}$$

where

$$\boldsymbol{\Gamma} = (\gamma_{ij})_{i,j=1}^d \geq \mathbf{0}, \ \boldsymbol{S}^+, \boldsymbol{S}^- \geq \mathbf{0}, \ \boldsymbol{L}_1, \boldsymbol{L}_2 \succeq \mathbf{0}, \quad \text{and} \quad \theta_0 \in \mathbb{R}$$

are the corresponding dual variables for the constraints. The augmentation helps us to classify the functional form of p as an exponential family distribution. We will derive this form from the Lagrangian

$$\begin{aligned}\mathcal{L} &= -H(p) + \theta_0\Big(\sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x}) - 1\Big) + \sum_{i,j} \gamma_{ij}(\|\boldsymbol{T}_{ij}\|_2^2 - c^2) \\ &\qquad - \big\langle \boldsymbol{S}^+ + \boldsymbol{S}^-, \boldsymbol{T}\big\rangle + \langle \boldsymbol{S}^+ - \boldsymbol{S}^- + \boldsymbol{L}_1 - \boldsymbol{L}_2, \hat{\boldsymbol{\Sigma}} - \mathbb{E}[\Sigma]\rangle - \lambda\,\langle \boldsymbol{L}_1 + \boldsymbol{L}_2, \boldsymbol{I}\rangle \\ &= -H(p) + \theta_0\Big(\sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x}) - 1\Big) + \sum_{i,j} \gamma_{ij}(\|\boldsymbol{T}_{ij}\|_2^2 - c^2) - \lambda\,\mathrm{tr}(\boldsymbol{L}_1 + \boldsymbol{L}_2) \\ &\qquad - \langle |\boldsymbol{S}|, \boldsymbol{T}\rangle + \langle \boldsymbol{S} + \boldsymbol{L}_1 - \boldsymbol{L}_2, \hat{\boldsymbol{\Sigma}} - \mathbb{E}[\Sigma]\rangle. \end{aligned} \tag{C.3}$$

For the second equation we wrote $\boldsymbol{S} = \boldsymbol{S}^+ - \boldsymbol{S}^-$ and $|\boldsymbol{S}| = \boldsymbol{S}^+ + \boldsymbol{S}^-$. The latter is possible because in a solution for each pair of corresponding entries of $\boldsymbol{S}^+$ and $\boldsymbol{S}^-$ at least one must be zero, respectively. Finally we used $\langle \boldsymbol{L}_1 + \boldsymbol{L}_2, \boldsymbol{I}\rangle = \mathrm{tr}(\boldsymbol{L}_1 + \boldsymbol{L}_2)$. Note that the Lagrangian is a function of the primal variables $\boldsymbol{T}$ and p, and of the dual variables.

In what follows, we set $\boldsymbol{\Theta} = \boldsymbol{S} + \boldsymbol{L}_1 - \boldsymbol{L}_2$. The discrete distribution p is given by the vector of probabilities $(p(\boldsymbol{x}))_{\boldsymbol{x}\in\mathcal{X}}$. The saddle point condition for the Lagrangian for $p(\boldsymbol{x})$ for fixed $\boldsymbol{x} \in \mathcal{X}$ implies that

$$\begin{aligned} 0 \overset{!}{=} \frac{\partial \mathcal{L}}{\partial p(\boldsymbol{x})} &= \frac{\partial}{\partial p(\boldsymbol{x})}\left(-H(p) + \theta_0 \sum_{\boldsymbol{x}'\in\mathcal{X}} p(\boldsymbol{x}') - \langle \boldsymbol{\Theta}, \mathbb{E}[\Sigma]\rangle\right) \\ &= \frac{\partial}{\partial p(\boldsymbol{x})}\left(\sum_{\boldsymbol{x}'\in\mathcal{X}} p(\boldsymbol{x}')\log(p(\boldsymbol{x}')) + \theta_0 \sum_{\boldsymbol{x}'\in\mathcal{X}} p(\boldsymbol{x}') - \Big\langle \boldsymbol{\Theta}, \sum_{\boldsymbol{x}'\in\mathcal{X}} p(\boldsymbol{x}')\Sigma(\boldsymbol{x}')\Big\rangle\right) \\ &= \log(p(\boldsymbol{x})) + 1 + \theta_0 - \langle \boldsymbol{\Theta}, \Sigma(\boldsymbol{x})\rangle \end{aligned}$$

Therefore, p must be of the form

$$p(\boldsymbol{x}) = \exp\left(\langle \boldsymbol{\Theta}, \Sigma(\boldsymbol{x})\rangle - a(\boldsymbol{\Theta})\right)$$

with normalization function $a(\boldsymbol{\Theta}) = \theta_0 + 1$. Now,

$$\begin{aligned} -H(p) = \mathbb{E}_p[\log p] &= \sum_{\boldsymbol{x}\in\mathcal{X}} p(\boldsymbol{x})[\langle \boldsymbol{\Theta}, \Sigma(\boldsymbol{x})\rangle - a(\boldsymbol{\Theta})] = \langle \boldsymbol{\Theta}, \mathbb{E}[\Sigma]\rangle - a(\boldsymbol{\Theta}) \\ &= \langle \boldsymbol{\Theta}, \mathbb{E}[\Sigma]\rangle + \ell(\boldsymbol{\Theta}) - \langle \boldsymbol{\Theta}, \hat{\boldsymbol{\Sigma}}\rangle = \ell(\boldsymbol{\Theta}) - \langle \boldsymbol{\Theta}, \hat{\boldsymbol{\Sigma}} - \mathbb{E}[\Sigma]\rangle, \end{aligned}$$

which follows since the log-likelihood is given by $\ell(\boldsymbol{\Theta}) = \langle \boldsymbol{\Theta}, \hat{\boldsymbol{\Sigma}}\rangle - a(\boldsymbol{\Theta})$, see Equation (3.5). We substitute p and $H(p)$ into the Lagrangian (C.3) to obtain

$$\begin{aligned} \mathcal{L} &= \ell(\boldsymbol{\Theta}) - \lambda\,\mathrm{tr}(\boldsymbol{L}_1 + \boldsymbol{L}_2) + \sum_{i,j} \gamma_{ij}(\|\boldsymbol{T}_{ij}\|_2^2 - c^2) - \langle |\boldsymbol{S}|, \boldsymbol{T}\rangle \\ &= \ell(\boldsymbol{\Theta}) - \lambda\,\mathrm{tr}(\boldsymbol{L}_1 + \boldsymbol{L}_2) + \sum_{i,j} \Big(\gamma_{ij}(\|\boldsymbol{T}_{ij}\|_2^2 - c^2) - \langle |\boldsymbol{S}_{ij}|, \boldsymbol{T}_{ij}\rangle\Big), \end{aligned}$$

recalling that $\boldsymbol{T}_{ij}$ and $\boldsymbol{S}_{ij}$ are the (i,j)-th sub-blocks of the matrices $\boldsymbol{T}$ and $\boldsymbol{S}$, respectively. Next, we minimize the Lagrangian in $\boldsymbol{T}$. The derivative w.r.t. $\boldsymbol{T}_{ij}$ is given by

$$\mathbf{0} \overset{!}{=} \frac{\partial \mathcal{L}}{\partial \boldsymbol{T}_{ij}} = 2\gamma_{ij}\boldsymbol{T}_{ij} - |\boldsymbol{S}_{ij}|.$$

Setting the gradient to zero yields $\boldsymbol{T}_{ij} = \frac{|\boldsymbol{S}_{ij}|}{2\gamma_{ij}}$ if $\gamma_{ij} > 0$. Otherwise, if $\gamma_{ij} = 0$ and $\boldsymbol{S}_{ij}$ has a non-zero entry, the Lagrangian is unbounded below. In any other case, all terms that include $\boldsymbol{T}_{ij}$ in the Lagrangian vanish. We substitute in the Lagrangian which gives us the dual function $g = g(\boldsymbol{\Gamma}, \boldsymbol{S}, \boldsymbol{L}_1, \boldsymbol{L}_2)$

$$g = \ell(\boldsymbol{\Theta}) - \lambda \operatorname{tr}(\boldsymbol{L}_1 + \boldsymbol{L}_2) + \sum_{i,j=1}^{d} \begin{cases} -c^2\gamma_{ij} - \frac{1}{4\gamma_{ij}}\|\boldsymbol{S}_{ij}\|_2^2, & \gamma_{ij} > 0, \\ 0, & \gamma_{ij} = 0, \boldsymbol{S}_{ij} \equiv \mathbf{0}, \\ -\infty, & \gamma_{ij} = 0, \boldsymbol{S}_{ij} \not\equiv \mathbf{0} \end{cases}$$

We eliminate the variables in $\boldsymbol{\Gamma}$ by finding an analytical solution for them. For $\gamma_{ij} > 0$ the gradient condition states

$$0 \overset{!}{=} \frac{\partial g}{\partial \gamma_{ij}} = -c^2 + \frac{1}{4\gamma_{ij}^2}\|\boldsymbol{S}_{ij}\|_2^2,$$

which is satisfied for $\gamma_{ij} = \frac{1}{2c}\|\boldsymbol{S}_{ij}\|_2$. With this γ_{ij} it holds $-c^2\gamma_{ij} - \frac{1}{4\gamma_{ij}}\|\boldsymbol{S}_{ij}\|_2^2 = -c\|\boldsymbol{S}_{ij}\|_2$. Note that this is also consistent with the second case above, so after the maximization w.r.t. $\boldsymbol{\Gamma}$ the dual function becomes

$$\begin{aligned} g(\boldsymbol{S}, \boldsymbol{L}_1, \boldsymbol{L}_2) &= \ell(\boldsymbol{\Theta}) - \lambda \operatorname{tr}(\boldsymbol{L}_1 + \boldsymbol{L}_2) - c\sum_{i,j} \|\boldsymbol{S}_{ij}\|_2 \\ &= \ell(\boldsymbol{S} + \boldsymbol{L}_1 - \boldsymbol{L}_2) - \lambda \operatorname{tr}(\boldsymbol{L}_1 + \boldsymbol{L}_2) - c\|\boldsymbol{S}\|_{1,2}, \end{aligned} \tag{C.4}$$

where $\boldsymbol{L}_1, \boldsymbol{L}_2 \succeq \mathbf{0}$. The dual function can be further simplified. To see this, for a given (symmetric) matrix $\boldsymbol{L} \in \operatorname{Sym}(m)$, consider the problem

$$\min_{\boldsymbol{L}_1, \boldsymbol{L}_2 \succeq \mathbf{0}} \quad \operatorname{tr}(\boldsymbol{L}_1 + \boldsymbol{L}_2) \quad \text{s.t.} \quad \boldsymbol{L} = \boldsymbol{L}_1 - \boldsymbol{L}_2. \tag{C.5}$$

Let $\boldsymbol{L} = \sum_{i=1}^{m} \sigma_i \boldsymbol{v}_i \boldsymbol{v}_i^\mathsf{T}$ be a singular decomposition of $\boldsymbol{L}$. Then, Problem (C.5) is minimized by $\boldsymbol{L}_1 = \sum_{i:\sigma_i \geq 0} \sigma_i \boldsymbol{v}_i \boldsymbol{v}_i^\mathsf{T}$ and $\boldsymbol{L}_2 = -\sum_{i:\sigma_i < 0} \sigma_i \boldsymbol{v}_i \boldsymbol{v}_i^\mathsf{T}$. Clearly, with this solution it holds $\operatorname{tr}(\boldsymbol{L}_1 + \boldsymbol{L}_2) = \|\boldsymbol{L}\|_* = \sum_{i=1}^{m} |\sigma_i|$ in the objective of Problem (C.5).

The solution of Problem (C.5) allows us to write $\boldsymbol{L} = \boldsymbol{L}_1 - \boldsymbol{L}_2$ in (C.4) and to replace the term $\operatorname{tr}(\boldsymbol{L}_1 + \boldsymbol{L}_2)$ by $\|\boldsymbol{L}\|_*$. Thereby, the variables $\boldsymbol{L}_1$ and $\boldsymbol{L}_2$ are eliminated. This leads to the final simplification of the dual function

$$g(\boldsymbol{S}, \boldsymbol{L}) = \ell(\boldsymbol{S} + \boldsymbol{L}) - \lambda\|\boldsymbol{L}\|_* - c\|\boldsymbol{S}\|_{1,2}.$$

Maximizing this dual function exactly matches the claimed regularized log-likelihood maximization problem. This finishes the proof. ■

If the constraint $\|\mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}}\| \leq \lambda$ in Problem (C.1) is replaced by the one-sided constraint $\hat{\boldsymbol{\Sigma}} - \mathbb{E}[\Sigma] \preceq \lambda \boldsymbol{I}$ with dual variable $\boldsymbol{L}_1$, then all terms related to the other side $\mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}} \preceq \lambda \boldsymbol{I}$ of the spectral norm constraint disappear. In particular, the corresponding dual variable $\boldsymbol{L}_2$ is removed from the dual function (C.4). Hence, in this case, the dual problem maximizes the objective function

$$g(\boldsymbol{S}, \boldsymbol{L}) = \ell(\boldsymbol{S} + \boldsymbol{L}) - \lambda \operatorname{tr}(\boldsymbol{L}) - c\|\boldsymbol{S}\|_{1,2} \quad \text{subject to} \quad \boldsymbol{L} = \boldsymbol{L}_1 \succeq \boldsymbol{0}.$$

C.1.2 Limited generalizability to continuous distributions

In this section, we undertake a small excursion and discuss entropy and relative entropy of continuous random variables. For simplicity, we limit the discussion to univariate positive distributions p on $\mathbb{R}$, that is, $p > 0$ everywhere. Then, the *differential entropy* of this distribution is defined as

$$h(p) = -\int_{\mathbb{R}} p(x) \log p(x)\, dx.$$

We first attempt to obtain differential entropy as the limit of discrete *Shannon* entropies by quantifying continuous distributions using discrete bins.

Entropy in the limit. Let p be a univariate distribution on $\mathbb{R}$ as above and let $\varepsilon > 0$. By the mean-value theorem, for $k \in \mathbb{Z}$, there exist $x_k = x_k(\varepsilon) \in [k\varepsilon, (k+1)\varepsilon)$ such that $\int_{k\varepsilon}^{(k+1)\varepsilon} p(x)\, dx = \varepsilon p(x_k)$. Define the discrete distributions p^ε with values x_k and corresponding probabilities $p^\varepsilon(x_k) = \varepsilon p(x_k)$ for $k \in \mathbb{Z}$. These distributions are normalized because

$$\sum_{k \in \mathbb{Z}} p^\varepsilon(x_k) = \sum_{k \in \mathbb{Z}} \varepsilon p(x_k) = \sum_{k \in \mathbb{Z}} \int_{k\varepsilon}^{(k+1)\varepsilon} p(x)\, dx = \int_{\mathbb{R}} p(x)\, dx = 1.$$

Now, the differential entropy of p can be obtained as the limit of Riemann sums, where for fixed $\varepsilon > 0$ we use the supporting points and intervals $x_k = x_k(\varepsilon) \in [k\varepsilon, (k+1)\varepsilon)$ from above (for $k \in \mathbb{Z}$). This gives us

$$\begin{aligned}
h(p) = -\int_{\mathbb{R}} p(x) \log p(x)\, dx &= \lim_{\varepsilon \to 0} -\sum_{k \in \mathbb{Z}} \varepsilon p(x_k) \log p(x_k) \\
&= \lim_{\varepsilon \to 0} -\sum_{k \in \mathbb{Z}} p^\varepsilon(x_k) \log \frac{p^\varepsilon(x_k)}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \left(-\sum_{k \in \mathbb{Z}} p^\varepsilon(x_k) \log p^\varepsilon(x_k) + \sum_{k \in \mathbb{Z}} p^\varepsilon(x_k) \log \varepsilon \right) \\
&= \lim_{\varepsilon \to 0} \left(-\sum_{k \in \mathbb{Z}} p^\varepsilon(x_k) \log p^\varepsilon(x_k) + \log \varepsilon \right) \\
&= \lim_{\varepsilon \to 0} \left(H(p^\varepsilon) + \log \varepsilon \right),
\end{aligned}$$

where the fifth equality follows from the fact that the distribution p^ε is normalized for all $\varepsilon > 0$. Assuming that $h(p)$ is finite, the calculation shows that $H(p^\varepsilon)$ diverges to infinity in the limit. Hence, a straightforward generalization of the discrete Shannon entropy by using the limit of Shannon entropies of increasingly fine discretizations would imply that all (positive) continuous distributions have an entropy of infinity. This would not be an informative measure, though it intuitively makes sense since the uncertainty of a continuous distribution can be seen as infinite.

To sum up, differential entropy is not a straightforward generalization of Shannon entropy. Particularly, it cannot be interpreted in absolute terms as the uncertainty of a distribution. However, differential entropy can be interpreted relatively to the differential entropy of another distribution. This is because when approximating $h(p) - h(q)$ via Riemann sums and discrete distributions as above, then the $\log \varepsilon$ term cancels out. Hence, it makes sense to define

$$h(p) - h(q) = \lim_{\varepsilon \to 0} \left(H(p^\varepsilon) - H(q^\varepsilon) \right).$$

This means that the direct comparison of the differential entropies of different distributions allows us to rigorously say that one or the other distribution is more chaotic (has more 'uncertainty'). Of course, when estimating a density using the maximum entropy principle, it does not make a difference if the objective function is the differential entropy $h(p)$ or the difference $h(p) - h(q)$ because $h(q)$ is a constant term. However, from a philosophical point of view, it is not clear what a reasonable choice for the distribution q is, which one may find unsatisfying.

Nevertheless, despite the problems in the interpretation of differential entropy, it can be shown that the duality of maximum entropy and maximum likelihood still holds for continuous distributions when the differential entropy is used. We do not provide a full proof since it mostly follows the lines of the proof given above in Section C.1.1. Note that in this proof, we classified the functional form of the maximum-entropy distribution as an exponential family distribution by differentiating w.r.t. the parameters of the discrete distribution. In the case of continuous distributions, the maximum-entropy distributions are still exponential family distributions as can be shown using variational calculus, specifically the *Euler/Euler-Lagrange equation*, see [Fox, 1987, Theorem 1].

Relative Entropy in the limit. We have seen that the difference of the differential entropies of two different continuous distributions on the same domain yields an interpretable measure for the difference of the 'uncertainty' between these distributions. Here, to round up the discussion of continuous entropy, we also consider *relative entropy* (also called *Kullback-Leibler divergence*), which is a measure of distance between distributions. It is defined for two distributions p and q on $\mathbb{R}$ as follows

$$D(p, q) = \int_{\mathbb{R}} p(x) \log \frac{p(x)}{q(x)}\, dx = -h(p) + H(p, q),$$

where

$$H(p,q) = -\int_{\mathbb{R}} p(x)\log q(x)\,dx$$

is the *cross entropy.* For $\varepsilon > 0$, define $p^{\varepsilon}(x_k)$ as above. Note that the discrete density defined by $q^{\varepsilon}(x_k) = \varepsilon q(x_k)$ for $k \in \mathbb{Z}$ is not normalized. However, the mass of q^{ε} converges to one as $\varepsilon \to 0$ as can be seen by the following convergence of Riemann sums:

$$\lim_{\varepsilon\to 0}\sum_{k\,\in\,\mathbb{Z}} q^{\varepsilon}(x_k) = \lim_{\varepsilon\to 0}\sum_{k\,\in\,\mathbb{Z}} \varepsilon q(x_k) = \int_{\mathbb{R}} q(x)\,dx = 1.$$

Since ultimately we are interested in the behavior in the limit, we treat q^{ε} as a regular distribution. We now show that cross entropy behaves similarly as differential entropy when we try to represent it as a limit of cross entropies of discrete distributions. To see that, we make use of Riemann sums as before:

$$\begin{aligned}
D(p,q) &= -\int_{\mathbb{R}} p(x)\log q(x)\,dx \\
&= \lim_{\varepsilon\to 0} -\sum_{k\,\in\,\mathbb{Z}} \varepsilon p(x_k)\log q(x_k) \\
&= \lim_{\varepsilon\to 0} -\sum_{k\,\in\,\mathbb{Z}} p^{\varepsilon}(x_k)\log \frac{q^{\varepsilon}(x_k)}{\varepsilon} \\
&= \lim_{\varepsilon\to 0}\left(-\sum_{k\,\in\,\mathbb{Z}} p^{\varepsilon}(x_k)\log q^{\varepsilon}(x_k) + \sum_{k\,\in\,\mathbb{Z}} p^{\varepsilon}(x_k)\log\varepsilon\right) \\
&= \lim_{\varepsilon\to 0}\left(-\sum_{k\,\in\,\mathbb{Z}} p^{\varepsilon}(x_k)\log q^{\varepsilon}(x_k) + \log\varepsilon\right) \\
&= \lim_{\varepsilon\to 0}\left(H(p^{\varepsilon},q^{\varepsilon}) + \log\varepsilon\right).
\end{aligned}$$

Here, in the second-to-last equality we used that p^{ε} is a normalized distribution, that is, it holds that $\sum_{k\,\in\,\mathbb{Z}} p^{\varepsilon}(x_k) = 1$. As a consequence, it holds for the limit of discrete relative entropies that

$$\begin{aligned}
D(p^{\varepsilon},q^{\varepsilon}) &= \sum_{k\,\in\,\mathbb{Z}} p^{\varepsilon}(x_k)\log\frac{p^{\varepsilon}(x_k)}{q^{\varepsilon}(x_k)} = \sum_{k\,\in\,\mathbb{Z}} p^{\varepsilon}(x_k)\log p^{\varepsilon}(x_k) - \sum_{k\,\in\,\mathbb{Z}} p^{\varepsilon}(x_k)\log q^{\varepsilon}(x_k) \\
&= -H(p^{\varepsilon}) + H(p^{\varepsilon},q^{\varepsilon}) \\
&= \underbrace{-H(p^{\varepsilon}) - \log\varepsilon}_{\to -h(p),\,\varepsilon\to 0} + \underbrace{\log\varepsilon + H(p^{\varepsilon},q^{\varepsilon})}_{\to H(p,q),\,\varepsilon\,\to\,0} \overset{\varepsilon\to 0}{\to} -h(p) + H(p,q) = D(p,q)
\end{aligned}$$

That means, relative entropy can indeed be seen as the generalization of discrete relative entropy.

C.2 Fused Latent and Graphical Models: Details

Pairwise conditional Gaussian distributions. A pairwise conditional Gaussian distribution on the sample space $\mathcal{X} \times \mathcal{Y} = \prod_{i=1}^{d}\{0, \dots, m_i\} \times \mathbb{R}^q$ is given by

$$\begin{aligned}
p(\boldsymbol{x}, \boldsymbol{y}) &= \exp\Bigg\{ \frac{1}{2} \sum_{i,j=1}^{d} \sum_{k=1}^{m_i} \sum_{l=1}^{m_j} q_{ij;kl} \mathbb{1}[x_i = k]\, \mathbb{1}[x_j = l] \\
&\qquad \dots + \sum_{s=1}^{q} \sum_{i=1}^{d} \sum_{k=1}^{m_i} \rho_{si;k} \mathbb{1}[x_i = k] y_s - \frac{1}{2} \sum_{s,t=1}^{q} \lambda_{st} y_s y_t - a(\boldsymbol{\Theta}) \Bigg\} \\
&= \exp\left\{ \frac{1}{2} \overline{\boldsymbol{x}}^{\mathsf{T}} \boldsymbol{Q}\, \overline{\boldsymbol{x}} + \boldsymbol{y}^{\mathsf{T}} \boldsymbol{R}\, \overline{\boldsymbol{x}} - \frac{1}{2} \boldsymbol{y}^{\mathsf{T}} \boldsymbol{\Lambda} \boldsymbol{y} - a(\boldsymbol{\Theta}) \right\} \\
&= \exp\left\{ \frac{1}{2} (\overline{\boldsymbol{x}}, \boldsymbol{y})^{\mathsf{T}} \boldsymbol{\Theta} (\overline{\boldsymbol{x}}, \boldsymbol{y}) - a(\boldsymbol{\Theta}) \right\},
\end{aligned}$$

where we omitted indicators for the 0-th levels of the discrete variables to ensure a unique parameterization, and we summarized the pairwise interaction parameters in the symmetric matrix

$$\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{R}^{\mathsf{T}} \\ \boldsymbol{R} & -\boldsymbol{\Lambda} \end{pmatrix} \in \mathrm{Sym}(m+q) = \mathrm{Sym}(w).$$

Of course, pairwise densities can also be extended with additional univariate parameters. However, we leave them out for simplicity since our main interest lies in modeling pairwise interactions between the variables. Besides, the diagonal of the discrete-discrete interaction parameters in $\boldsymbol{Q}$ can be seen as univariate parameters for the discrete variables. Note the group structure of the matrix $\boldsymbol{\Theta}$:

- $\boldsymbol{Q} \in \mathrm{Sym}(m)$ contains the groups $\boldsymbol{Q}_{ij} = (q_{ij;kl})_{k\in[m_i], l\in[m_j]} \in \mathbb{R}^{m_i \times m_j}$ of discrete-discrete interactions for $i, j \in [d]$,
- $\boldsymbol{R} \in \mathbb{R}^{q\times m}$ contains the groups $\boldsymbol{r}_{si} = (\rho_{si;k})_{k\in[m_i]} \in \mathbb{R}^{m_i}$ of quantitative-discrete interactions for $i \in [d]$ and $s \in [q]$, and
- $\boldsymbol{0} \prec \boldsymbol{\Lambda} \in \mathrm{Sym}(q)$ contains the quantitative-quantitative interaction parameters λ_{st} for $s, t \in [q]$, that is, the groups of quantitative-quantitative interactions consist of only single elements.

In summary, the group structure reads as

$$\boldsymbol{\Theta} = \left(\begin{array}{c|c|c|c|c|c}
\boldsymbol{Q}_{11} & \cdots & \boldsymbol{Q}_{1d} & \boldsymbol{r}_{11}^{\mathsf{T}} & \cdots & \boldsymbol{r}_{q1}^{\mathsf{T}} \\ \hline
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \hline
\boldsymbol{Q}_{d1} & \cdots & \boldsymbol{Q}_{dd} & \boldsymbol{r}_{1d}^{\mathsf{T}} & \cdots & \boldsymbol{r}_{qd}^{\mathsf{T}} \\ \hline
\boldsymbol{r}_{11} & \cdots & \boldsymbol{r}_{1d} & -\lambda_{11} & \cdots & -\lambda_{1q} \\ \hline
\vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ \hline
\boldsymbol{r}_{q1} & \cdots & \boldsymbol{r}_{qd} & -\lambda_{q1} & \cdots & -\lambda_{qq}
\end{array} \right).$$

In our model definitions, the discrete variables always precede the quantitative variables. We denote the $(m_i \times m_j)$-sub-matrix of the interaction parameter matrix $\boldsymbol{\Theta}$ that describes the interaction between the i-th and j-th variable by $\boldsymbol{\Theta}_{ij}$. Here, $i, j \in [d+q]$ and we define $m_i = 1$ for $i = d+1, \ldots, d+q$ (the quantitative variables).

Marginalization of conditional Gaussian (CG) variables from a CG model. Based on a partition of the quantitative, conditional Gaussian variables into q observed and r latent variables, the joint model can be written as

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) &\propto \exp\left(\frac{1}{2}\left(\overline{\boldsymbol{x}}, \boldsymbol{y}, \boldsymbol{z}\right)^\mathsf{T} \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{R}_y^\mathsf{T} & -\boldsymbol{R}_z^\mathsf{T} \\ \boldsymbol{R}_y & -\boldsymbol{\Lambda}_y & -\boldsymbol{\Lambda}_{zy}^\mathsf{T} \\ \boldsymbol{R}_z & -\boldsymbol{\Lambda}_{zy} & -\boldsymbol{\Lambda}_z \end{pmatrix} \left(\overline{\boldsymbol{x}}, \boldsymbol{y}, \boldsymbol{z}\right)\right), \\ &= \exp\left(\frac{1}{2}\left(\overline{\boldsymbol{x}}, \boldsymbol{y}\right)^\mathsf{T} \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{R}_y^\mathsf{T} \\ \boldsymbol{R}_y & -\boldsymbol{\Lambda}_y \end{pmatrix} \left(\overline{\boldsymbol{x}}, \boldsymbol{y}\right) + \boldsymbol{z}^\mathsf{T}\begin{pmatrix} \boldsymbol{R}_z & -\boldsymbol{\Lambda}_{zy} \end{pmatrix}\left(\overline{\boldsymbol{x}}, \boldsymbol{y}\right) - \frac{1}{2}\boldsymbol{z}^\mathsf{T}\boldsymbol{\Lambda}_z\boldsymbol{z}\right), \end{aligned}$$

where $(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \in \mathcal{X} \times \mathcal{Y} \times \mathcal{Z} = \prod_{i=1}^d \{0, \ldots, m_i\} \times \mathbb{R}^q \times \mathbb{R}^r$. Here, $\begin{pmatrix} \boldsymbol{Q} & \boldsymbol{R}_y^\mathsf{T} \\ \boldsymbol{R}_y & -\boldsymbol{\Lambda}_y \end{pmatrix}$ are the interaction parameters for the observed variables with discrete-discrete interaction parameters in $\boldsymbol{Q} \in \operatorname{Sym}(m)$, discrete-quantitative interactions in $\boldsymbol{R}_y \in \mathbb{R}^{q\times m}$, and quantitative-quantitative interactions in $\boldsymbol{\Lambda}_y \in \operatorname{Sym}(q)$. Moreover, $\begin{pmatrix} \boldsymbol{R}_z & -\boldsymbol{\Lambda}_{zy} \end{pmatrix} \in \mathbb{R}^{r\times(m+q)}$ are the interaction parameters for the quantitative latent variables with the observed variables, where specifically $\boldsymbol{R}_z \in \mathbb{R}^{r\times m}$ models the interactions with the observed discrete variables, and $\boldsymbol{\Lambda}_{zy} \in \mathbb{R}^{r\times q}$ models the interactions with the observed quantitative variables. Finally, $\boldsymbol{\Lambda}_z \in \operatorname{Sym}(r)$ with $\boldsymbol{\Lambda}_z \succ \boldsymbol{0}$ is the precision matrix of the quantitative latent variables. The density can also be written as

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \propto \exp\Bigg(&\frac{1}{2}(\overline{\boldsymbol{x}}, \boldsymbol{y})^\mathsf{T}\left[\begin{pmatrix} \boldsymbol{Q} & \boldsymbol{R}_y^\mathsf{T} \\ \boldsymbol{R}_y & -\boldsymbol{\Lambda}_y \end{pmatrix} + \begin{pmatrix} \boldsymbol{R}_z & -\boldsymbol{\Lambda}_{zy} \end{pmatrix}^\mathsf{T}\boldsymbol{\Lambda}_z^{-1}\begin{pmatrix} \boldsymbol{R}_z & -\boldsymbol{\Lambda}_{zy} \end{pmatrix}\right](\overline{\boldsymbol{x}}, \boldsymbol{y}) \\ &\ldots - \frac{1}{2}\left[\boldsymbol{z} - \boldsymbol{\Lambda}_z^{-1}\begin{pmatrix} \boldsymbol{R}_z & -\boldsymbol{\Lambda}_{zy} \end{pmatrix}(\overline{\boldsymbol{x}}, \boldsymbol{y})\right]^\mathsf{T}\boldsymbol{\Lambda}_z\left[\boldsymbol{z} - \boldsymbol{\Lambda}_z^{-1}\begin{pmatrix} \boldsymbol{R}_z & -\boldsymbol{\Lambda}_{zy} \end{pmatrix}(\overline{\boldsymbol{x}}, \boldsymbol{y})\right]\Bigg). \end{aligned}$$

Observe that for fixed values of $(\boldsymbol{x}, \boldsymbol{y})$ this is the unnormalized density of a multivariate Gaussian in $\boldsymbol{z}$ with mean vector $\boldsymbol{\Lambda}_z^{-1}\begin{pmatrix} \boldsymbol{R}_z & -\boldsymbol{\Lambda}_{zy} \end{pmatrix}(\overline{\boldsymbol{x}}, \boldsymbol{y})$ and precision matrix $\boldsymbol{\Lambda}_z$. Hence, the marginal distribution is

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}) &= \int_{\mathbb{R}^r} p(\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z})\, dz \\ &\propto \exp\left(\frac{1}{2}\left(\overline{\boldsymbol{x}}, \boldsymbol{y}\right)^\mathsf{T}\left[\begin{pmatrix} \boldsymbol{Q} & \boldsymbol{R}_y^\mathsf{T} \\ \boldsymbol{R}_y & -\boldsymbol{\Lambda}_y \end{pmatrix} + \begin{pmatrix} \boldsymbol{R}_z & -\boldsymbol{\Lambda}_{zy} \end{pmatrix}^\mathsf{T}\boldsymbol{\Lambda}_z^{-1}\begin{pmatrix} \boldsymbol{R}_z & -\boldsymbol{\Lambda}_{zy} \end{pmatrix}\right]\left(\overline{\boldsymbol{x}}, \boldsymbol{y}\right)\right), \end{aligned}$$

where we set $\boldsymbol{S} = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{R}_y^\mathsf{T} \\ \boldsymbol{R}_y & -\boldsymbol{\Lambda}_y \end{pmatrix}$ and $\boldsymbol{L} = \begin{pmatrix} \boldsymbol{R}_z & -\boldsymbol{\Lambda}_{zy} \end{pmatrix}^\mathsf{T}\boldsymbol{\Lambda}_z^{-1}\begin{pmatrix} \boldsymbol{R}_z & -\boldsymbol{\Lambda}_{zy} \end{pmatrix} \succeq \boldsymbol{0}$, which has at most rank r.

Normalization of a pairwise CG density. For $\boldsymbol{\Theta} = \begin{pmatrix} \boldsymbol{Q} & \boldsymbol{R}^{\mathsf{T}} \\ \boldsymbol{R} & -\boldsymbol{\Lambda} \end{pmatrix}$ we want to compute $a(\boldsymbol{\Theta})$ such that

$$\begin{aligned} p(\boldsymbol{x}, \boldsymbol{y}) &= \exp\left(\frac{1}{2}\,(\overline{\boldsymbol{x}}, \boldsymbol{y})^{\mathsf{T}} \boldsymbol{\Theta} (\overline{\boldsymbol{x}}, \boldsymbol{y}) - a(\boldsymbol{\Theta})\right) \\ &= \exp\left(\frac{1}{2}\,\overline{\boldsymbol{x}}^{\mathsf{T}} \boldsymbol{Q}\,\overline{\boldsymbol{x}} + \boldsymbol{y}^{\mathsf{T}} \boldsymbol{R}\,\overline{\boldsymbol{x}} - \frac{1}{2} \boldsymbol{y}^{\mathsf{T}} \boldsymbol{\Lambda} \boldsymbol{y} - a(\boldsymbol{\Theta})\right) \end{aligned}$$

is a normalized density. We get

$$\begin{aligned} a(\boldsymbol{\Theta}) &= \log\left(\sum_{\boldsymbol{x} \in \mathcal{X}} \int_{\mathcal{Y}} \exp\left(\frac{1}{2}\,\overline{\boldsymbol{x}}^{\mathsf{T}} \boldsymbol{Q}\,\overline{\boldsymbol{x}} + \boldsymbol{y}^{\mathsf{T}} \boldsymbol{R}\,\overline{\boldsymbol{x}} - \frac{1}{2} \boldsymbol{y}^{\mathsf{T}} \boldsymbol{\Lambda} \boldsymbol{y}\right)\, dy\right) \\ &= \log\left(\sum_{\boldsymbol{x} \in \mathcal{X}} \int_{\mathcal{Y}} \exp\left(\frac{1}{2} \overline{\boldsymbol{x}}^{\mathsf{T}} (\boldsymbol{Q} + \boldsymbol{R}^{\mathsf{T}} \boldsymbol{\Lambda}^{-1} \boldsymbol{R})\, \overline{\boldsymbol{x}} \right.\right. \\ &\qquad\qquad \left.\left. - \frac{1}{2} (\boldsymbol{y} - \boldsymbol{\Lambda}^{-1} \boldsymbol{R}\,\overline{\boldsymbol{x}})^{\mathsf{T}} \boldsymbol{\Lambda} (\boldsymbol{y} - \boldsymbol{\Lambda}^{-1} \boldsymbol{R}\,\overline{\boldsymbol{x}})\right) dy\right) \\ &= \log\left((2\pi)^{q/2} \det(\boldsymbol{\Lambda})^{-1/2} \sum_{\boldsymbol{x} \in \mathcal{X}} \exp\left(\frac{1}{2}\,\overline{\boldsymbol{x}}^{\mathsf{T}} (\boldsymbol{Q} + \boldsymbol{R}^{\mathsf{T}} \boldsymbol{\Lambda}^{-1} \boldsymbol{R})\, \overline{\boldsymbol{x}}\right)\right) \\ &= \frac{q}{2} \log(2\pi) - \frac{1}{2} \log\det \boldsymbol{\Lambda} + \log\left(\sum_{\boldsymbol{x} \in \mathcal{X}} \exp\left(\frac{1}{2}\,\overline{\boldsymbol{x}}^{\mathsf{T}} (\boldsymbol{Q} + \boldsymbol{R}^{\mathsf{T}} \boldsymbol{\Lambda}^{-1} \boldsymbol{R})\, \overline{\boldsymbol{x}}\right)\right). \end{aligned}$$

Likelihood of a pairwise CG density. Given observations $(\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)}) \in \mathcal{X} \times \mathbb{R}^q$ for $k = 1, \ldots, n$, the *negative* log-likelihood for pairwise parameters $\boldsymbol{\Theta}$ is given as

$$\begin{aligned} \ell(\boldsymbol{\Theta}) &= -\sum_{k=1}^{n} \log p(\boldsymbol{x}^{(k)}, \boldsymbol{y}^{(k)}) = -\sum_{k=1}^{n} \left(\frac{1}{2}\,(\overline{\boldsymbol{x}}^{(k)}, \boldsymbol{y}^{(k)})^{\mathsf{T}} \boldsymbol{\Theta} (\overline{\boldsymbol{x}}^{(k)}, \boldsymbol{y}^{(k)}) - a(\boldsymbol{\Theta})\right) \\ &= n\, a(\boldsymbol{\Theta}) - \frac{1}{2} \sum_{k=1}^{n} \left\langle (\overline{\boldsymbol{x}}^{(k)}, \boldsymbol{y}^{(k)}) (\overline{\boldsymbol{x}}^{(k)}, \boldsymbol{y}^{(k)})^{\mathsf{T}}, \boldsymbol{\Theta} \right\rangle = n\, a(\boldsymbol{\Theta}) - \frac{1}{2} \langle n \hat{\boldsymbol{\Sigma}}, \boldsymbol{\Theta} \rangle, \end{aligned}$$

where $\hat{\boldsymbol{\Sigma}} = 1/n \sum_{k=1}^{n} (\overline{\boldsymbol{x}}^{(k)}, \boldsymbol{y}^{(k)}) (\overline{\boldsymbol{x}}^{(k)}, \boldsymbol{y}^{(k)})^{\mathsf{T}}$ is the empirical second-moment matrix. Often, one additionally scales the (negative) log-likelihood with a factor $2/n$. In this thesis, we thus use with a slight abuse of notation

$$\ell(\boldsymbol{\Theta}) = 2a(\boldsymbol{\Theta}) - \langle \hat{\boldsymbol{\Sigma}}, \boldsymbol{\Theta} \rangle,$$

unless stated otherwise.

Gradient. Finally, in preparation of the theoretical analysis, we calculate the gradient of the log-partition function/negative log-likelihood function. Let

$$\hat{p}(\boldsymbol{x}, \boldsymbol{y}) = \exp\left(1/2\, (\overline{\boldsymbol{x}}, \boldsymbol{y})^{\mathsf{T}} \boldsymbol{\Theta} (\overline{\boldsymbol{x}}, \boldsymbol{y})\right) = \exp\left(1/2\, \langle \boldsymbol{\Theta}, (\overline{\boldsymbol{x}}, \boldsymbol{y})(\overline{\boldsymbol{x}}, \boldsymbol{y})^{\mathsf{T}} \rangle\right).$$

Let Θ_{ij} be any entry of $\boldsymbol{\Theta}$. Then,

$$\begin{aligned}
\frac{\partial a(\boldsymbol{\Theta})}{\partial \Theta_{ij}} &= \frac{\partial}{\partial \Theta_{ij}} \left(\log \left(\sum_{\boldsymbol{x} \in \mathcal{X}} \int_{\mathcal{Y}} \hat{p}(\boldsymbol{x}, \boldsymbol{y}) \, dy \right) \right) \\
&= \frac{\sum_{\boldsymbol{x} \in \mathcal{X}} \int_{\mathcal{Y}} 1/2 \, [(\overline{\boldsymbol{x}}, \boldsymbol{y})(\overline{\boldsymbol{x}}, \boldsymbol{y})^{\mathsf{T}}]_{ij} \hat{p}(\boldsymbol{x}, \boldsymbol{y}) \, dy}{\sum_{\boldsymbol{x} \in \mathcal{X}} \int_{\mathcal{Y}} \hat{p}(\boldsymbol{x}, \boldsymbol{y}) \, dy} \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} \int_{\mathcal{Y}} \frac{1}{2} [(\overline{\boldsymbol{x}}, \boldsymbol{y})(\overline{\boldsymbol{x}}, \boldsymbol{y})^{\mathsf{T}}]_{ij} \hat{p}(\boldsymbol{x}, \boldsymbol{y}) \exp(-a(\boldsymbol{\Theta})) \, dy \\
&= \sum_{\boldsymbol{x} \in \mathcal{X}} \int_{\mathcal{Y}} \frac{1}{2} [(\overline{\boldsymbol{x}}, \boldsymbol{y})(\overline{\boldsymbol{x}}, \boldsymbol{y})^{\mathsf{T}}]_{ij} p(\boldsymbol{x}, \boldsymbol{y}) \, dy = \frac{1}{2} \mathbb{E}[[(\overline{\boldsymbol{x}}, \boldsymbol{y})(\overline{\boldsymbol{x}}, \boldsymbol{y})^{\mathsf{T}}]_{ij}]
\end{aligned}$$

Hence, $\nabla a(\boldsymbol{\Theta}) = 1/2 \, \mathbb{E}[\Sigma]$, where $\Sigma = (\overline{\boldsymbol{x}}, \boldsymbol{y})(\overline{\boldsymbol{x}}, \boldsymbol{y})^{\mathsf{T}}$ and the expectation is w.r.t. the model parameterized with $\boldsymbol{\Theta}$. Consequently, the gradient of the negative log-likelihood is given by the sampling error $\nabla \ell(\boldsymbol{\Theta}) = \mathbb{E}[\Sigma] - \hat{\boldsymbol{\Sigma}}$ of the second-moment matrix.

C.3 Proof of Consistency

C.3.1 Outline for the proof of Theorem 3.3

Here, we sketch the proof of Theorem 3.3, which generalizes the ones in [Chandrasekaran et al., 2012; Nussbaum and Giesen, 2019b, 2020b] and reconciles a version of the primal-dual witness proof technique. The basic idea for the proof is to study the optimality conditions of the simplified problem

$$(\boldsymbol{S}_{\varnothing}, \boldsymbol{L}_{\varnothing}) = \operatorname*{arg\,min}_{\boldsymbol{S}, \boldsymbol{L}} \; \ell(\boldsymbol{S} + \boldsymbol{L}) + \lambda \left(\gamma \|\boldsymbol{S}\|_{1,2} + \|\boldsymbol{L}\|_{*} \right) \quad \text{s.t.} \quad \Lambda[\boldsymbol{S} + \boldsymbol{L}] \succ \boldsymbol{0}, \quad \text{(C.6)}$$

where we drop the constraint $\boldsymbol{L} \succeq \boldsymbol{0}$ from Problem (3.7) and subsequently write $\|\boldsymbol{L}\|_{*}$ for the nuclear norm. It turns out that under our assumptions the solution will automatically satisfy the dropped constraint. The primal-dual witness technique proceeds by constructing a primal-dual pair that satisfies the optimality conditions of Problem (C.6)

$$\begin{aligned}
\boldsymbol{0} &= \nabla \ell(\boldsymbol{S}_{\varnothing} + \boldsymbol{L}_{\varnothing}) + \boldsymbol{Z}_{1,2}, \qquad \boldsymbol{Z}_{1,2} \in \lambda \gamma \partial \|\boldsymbol{S}_{\varnothing}\|_{1,2} \\
\boldsymbol{0} &= \nabla \ell(\boldsymbol{S}_{\varnothing} + \boldsymbol{L}_{\varnothing}) + \boldsymbol{Z}_{*}, \qquad \boldsymbol{Z}_{*} \in \lambda \partial \|\boldsymbol{L}_{\varnothing}\|_{*}.
\end{aligned}$$

We call $(\boldsymbol{S}_{\varnothing}, \boldsymbol{L}_{\varnothing})$ a primal solution and the subgradients $(\boldsymbol{Z}_{1,2}, \boldsymbol{Z}_{*})$ dual solutions. We summarize these dual variables as $\boldsymbol{Z} = -(\boldsymbol{Z}_{1,2}, \boldsymbol{Z}_{*})$ which allows writing the optimality condition for Problem (C.6) compactly as

$$\boldsymbol{Z} = D \nabla \ell(\boldsymbol{S}_{\varnothing} + \boldsymbol{L}_{\varnothing}),$$

where the duplication operator D has been defined in Proposition 3.2. We will now outline the general proof strategy that consists of the following three steps. First, it is shown that the solution to the problem restricted to a certain *correct model set* is consistent. Second, it is proven that the solution remains unchanged when the correct model set is linearized. Third, it is verified that the solution to the linearized problem also solves the original Problem (3.7). With some more details these three steps are as follows:

(1) A version of Problem (C.6) that is additionally restricted to the correct model set $\mathcal{M}$ is considered, where $\mathcal{M}$ is constrained such that for $(\boldsymbol{S}, \boldsymbol{L}) \in \mathcal{M}$ the errors $\boldsymbol{S} - \boldsymbol{S}^\star$ and $\boldsymbol{L} - \boldsymbol{L}^\star$ are small in some sense, the group support of $\boldsymbol{S}$ is contained in the true group support of $\boldsymbol{S}^\star$, and the rank of $\boldsymbol{L}$ cannot be greater than the true rank of $\boldsymbol{L}^\star$. The last constraint turns the problem into a non-convex one. However, under the three main assumptions, these constraints entail consistency properties for all elements in $\mathcal{M}$, in particular for any solution $(\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}})$ to the (non-convex) problem restricted to $\mathcal{M}$. Specifically, for any $(\boldsymbol{S}, \boldsymbol{L}) \in \mathcal{M}$ it is shown that $\boldsymbol{S}$ has the correct group support and that $\boldsymbol{L}$ has the correct rank. Moreover, it it is shown that $\boldsymbol{L} \succeq \boldsymbol{0}$. Hence, any $(\boldsymbol{S}, \boldsymbol{L}) \in \mathcal{M}$ is feasible for the original Problem (3.7).

(2) A solution $(\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}})$ from the previous step is fixed. Let $\mathcal{Q}(\boldsymbol{S}_{\mathcal{M}})$ and $\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})$ be the respective tangent spaces to the group-sparse and low-rank matrix varieties at the solution $(\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}})$. Then, a *linearized* problem, where the constraint set $\mathcal{M}$ is replaced by the new constraint set $\mathcal{J} = \mathcal{Q}(\boldsymbol{S}_{\mathcal{M}}) \times \mathcal{T}(\boldsymbol{L}_{\mathcal{M}})$, is considered. This particularly replaces the non-convex rank constraint by an appropriate linear tangent-space constraint. It is shown that the solution $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ to the convex linearized problem is unique and satisfies all constraints from $\mathcal{M}$. In fact, this implies that the solution $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ equals $(\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}})$. Hence, it inherits the consistency properties from the first step.

(3) It is shown that $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ solves Problem (C.6) and qualifies as a *primal-dual witness*, that is, $\boldsymbol{Z} = D\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}})$ is shown to be *strictly* dual feasible and thus is a valid subgradient for the optimality condition of Problem (C.6). For showing this, note that the primal solution $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ from Step (2) to the problem restricted to $\mathcal{J}$ is already characterized by satisfying the optimality condition restricted to the components in $\mathcal{J}$. This is because additional Lagrange multipliers due to the tangent-space constraints lie in $\mathcal{J}^\perp$. Hence, to verify (strict) dual feasibility, it only remains to show that the optimality condition restricted to the components in $\mathcal{J}^\perp$ is also satisfied.

    Finally, using the primal-dual witness $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}}, \boldsymbol{Z})$ it is shown that *all* primal solutions to Problem (C.6) must be in $\mathcal{J}$. From that it follows that $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$, which has consistency properties from the previous steps and is the unique solution in $\mathcal{J}$, must also be the unique solution to Problem (C.6). Since $\boldsymbol{L}_{\mathcal{M}} = \boldsymbol{L}_{\mathcal{J}} \succeq \boldsymbol{0}$, it can be concluded that $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ also solves Problem (3.7).

The proof given in the following sections follows precisely these three steps.

C.3.2 Constants

Here, we give an overview of constants that are used in the proof or are necessary to refine the problem-specific constants that appear in the assumptions and claims of Theorem 3.3.

First, we frequently encounter the constant

$$\chi = \eta \max\left\{\frac{\nu\alpha}{3\beta(2-\nu)}, 1\right\},$$

which by Lemma C.4 in Section C.3.4 essentially is a norm compatibility constant between the γ-norm (for any $\gamma \in [\gamma_{\min}, \gamma_{\max}]$) and the spectral norm. Besides this norm compatibility constant, further problem-specific constants appear in the proof. First, let $r_0 > 0$ be such that all $\boldsymbol{\Theta}$ in the spectral-norm ball with radius r_0 around $\boldsymbol{\Theta}^\star = \boldsymbol{S}^\star + \boldsymbol{L}^\star$ are feasible in the sense that the pairwise CG density parametrized by $\boldsymbol{\Theta}$ is normalizable, that is, $\Lambda[\boldsymbol{\Theta}] \succ \boldsymbol{0}$. Such a r_0 exists because the feasible domain is an open set. If there are no quantitative variables, any r_0 can be chosen. Second, $l(r_0) > 0$ is a Lipschitz constant of the Hessian $\nabla^2 \ell = \nabla^2 a$ on a ball with radius r_0 around $\boldsymbol{\Theta}^\star$, see Lemma C.11. Additionally, the following problem-specific constants appear in the proof:

$$c_0 = 2l(r_0)\chi \max\left\{1, \frac{\nu\alpha}{2\beta(2-\nu)}\right\}^2, \qquad c_1 = \max\left\{1, \frac{\nu\alpha}{2\beta(2-\nu)}\right\}^{-1} \frac{r_0}{2},$$

$$c_2 = \frac{40}{\alpha} + \frac{1}{\|H^\star\|}, \qquad c_3 = \left(\frac{6(2-\nu)}{\nu} + 1\right) c_2^2 \|H^\star\| \chi,$$

$$c_4 = c_2 + \frac{3\alpha c_2^2(2-\nu)}{16(3-\nu)}, \qquad c_5 = \frac{\nu\alpha c_2}{2\beta(2-\nu)},$$

where

$$\|H^\star\| = \max_{\boldsymbol{M} \in \mathrm{Sym}(w) \colon \|\boldsymbol{M}\|=1} \|H^\star \boldsymbol{M}\|$$

is the operator norm of the Hessian $H^\star = \nabla^2 \ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star)$. Moreover, α and δ are defined in Assumption 1, $\nu = (1-\delta/\alpha)/2$, and β is defined in Assumption 2. Next, the precise constants for the gap Assumption 3 are given as follows:

$$s_{\min} > \frac{c_5 \lambda}{\mu(\mathcal{Q})} \qquad \text{and} \qquad \sigma_{\min} \geq \max\left\{\frac{c_3\eta}{\xi(\mathcal{T})^2}, c_4\right\} \lambda.$$

The precise definitions of all remaining constants that appear in Theorem 3.3 and the following discussion are given as follows:

$$C_1 = \frac{3\alpha(2-\nu)}{32(3-\nu)} c_1, \qquad C_2 = \frac{3\alpha^2\nu(2-\nu)}{2^{11} c_0 (3-\nu)^2},$$

$$C_3 = \frac{\nu}{6(2-\nu)}, \qquad C_4 = \frac{32(3-\nu)}{3\alpha(2-\nu)}, \qquad C_5 = \frac{6(2-\nu)\chi}{\nu}.$$

Finally, we derive a simple bound on $\xi(\mathcal{T}(\boldsymbol{L}))$. For that, let $\eta = \max_{i \in [d]} m_i$ such that η^2 is an upper bound for the number of elements in a group of interaction parameters. Then, we have the general norm bounds $\|\cdot\|_{\infty,2} \le \eta\|\cdot\|_\infty \le \eta\|\cdot\|$, which imply that $\xi(\mathcal{T}(\boldsymbol{L})) \le \eta$.

C.3.3 Proof of Proposition 3.2: Coupled stability

In preparation for the proof of Proposition 3.2, we show a few auxiliary results. The first lemma is important for the whole proof and is an adaptation of the projection Lemma B.2 (and hence can be proven similarly).

Lemma C.1. *For any two tangent spaces $\mathcal{Q}(\boldsymbol{S})$ at smooth point $\boldsymbol{S} \in \mathcal{S}(s)$ and $\mathcal{T}(\boldsymbol{L})$ at smooth point $\boldsymbol{L} \in \mathcal{L}(r)$ we can bound the norms of projections of matrices $\boldsymbol{M}, \boldsymbol{N} \in \mathrm{Sym}(w)$ in the following manner:*

$$\|P_{\mathcal{Q}(\boldsymbol{S})}\boldsymbol{M}\|_{\infty,2} \le \|\boldsymbol{M}\|_{\infty,2} \quad \textit{and} \quad \|P_{\mathcal{Q}(\boldsymbol{S})^\perp}\boldsymbol{M}\|_{\infty,2} \le \|\boldsymbol{M}\|_{\infty,2}$$
$$\|P_{\mathcal{T}(\boldsymbol{L})}\boldsymbol{N}\| \le 2\|\boldsymbol{N}\| \quad \textit{and} \quad \|P_{\mathcal{T}(\boldsymbol{L})^\perp}\boldsymbol{N}\| \le \|\boldsymbol{N}\|.$$

In particular, for $\mathcal{J} = \mathcal{Q}(\boldsymbol{S}) \times \mathcal{T}(\boldsymbol{L})$ we have

$$\|P_{\mathcal{J}}(\boldsymbol{M},\boldsymbol{N})\|_\gamma \le 2\,\|(\boldsymbol{M},\boldsymbol{N})\|_\gamma \quad \textit{and} \quad \|P_{\mathcal{J}^\perp}(\boldsymbol{M},\boldsymbol{N})\|_\gamma \le \|(\boldsymbol{M},\boldsymbol{N})\|_\gamma\,.$$

The next auxiliary lemma can be used to bound the norm-compatibility constant ξ for a low-rank tangent space by the one for a nearby low-rank tangent space.

Lemma C.2. *Let $\mathcal{T}_1, \mathcal{T}_2 \subseteq \mathrm{Sym}(w)$ be two matrix subspaces of the same dimension with bounded twisting in the sense that*

$$\rho(\mathcal{T}_1,\mathcal{T}_2) = \max_{\|\boldsymbol{M}\|=1} \|[P_{\mathcal{T}_1} - P_{\mathcal{T}_2}](\boldsymbol{M})\| < 1.$$

Then, it holds that

$$\xi(\mathcal{T}_2) \le \frac{\xi(\mathcal{T}_1) + \eta\rho(\mathcal{T}_1,\mathcal{T}_2)}{1-\rho(\mathcal{T}_1,\mathcal{T}_2)}.$$

Proof. The proof follows the lines of [Chandrasekaran et al., 2012, Lemma 3.1]. First note that the projection $\mathcal{T}_1 \to \mathcal{T}_2, \boldsymbol{M} \mapsto P_{\mathcal{T}_2}\boldsymbol{M}$ is bijective since $\mathcal{T}_1$ and $\mathcal{T}_2$ have the same dimension and the projection is injective since for any $\boldsymbol{0} \ne \boldsymbol{M} \in \mathcal{T}_1$ it holds

$$\begin{aligned}\|P_{\mathcal{T}_2}\boldsymbol{M}\| &= \|P_{\mathcal{T}_1}\boldsymbol{M} + (P_{\mathcal{T}_2} - P_{\mathcal{T}_1})\boldsymbol{M}\|\\ &\ge \|P_{\mathcal{T}_1}\boldsymbol{M}\| - \|(P_{\mathcal{T}_2} - P_{\mathcal{T}_1})\boldsymbol{M}\|\\ &\ge \|\boldsymbol{M}\| - \rho(\mathcal{T}_1,\mathcal{T}_2)\|\boldsymbol{M}\| = (1-\rho(\mathcal{T}_1,\mathcal{T}_2))\|\boldsymbol{M}\| > 0,\end{aligned}$$

where the first inequality is the triangle inequality, the second inequality uses the definition of the twisting ρ, and the last inequality follows from the assumption $\rho(\mathcal{T}_1,\mathcal{T}_2) < 1$. The calculation also implies that the ball $\{\boldsymbol{M} \in \mathcal{T}_2 : \|\boldsymbol{M}\| \le 1\}$ is contained in the image of the ball $\{\boldsymbol{M} \in \mathcal{T}_1 : \|\boldsymbol{M}\| \le 1/(1-\rho(\mathcal{T}_1, \mathcal{T}_2))\}$ under $P_{\mathcal{T}_2}$. Hence, we have that

$$\begin{aligned}
\xi(\mathcal{T}_2) &= \max_{\boldsymbol{M}\in\mathcal{T}_2,\, \|\boldsymbol{M}\|\le 1} \|\boldsymbol{M}\|_{2,\infty} \\
&\le \max_{\boldsymbol{M}\in\mathcal{T}_1,\, \|\boldsymbol{M}\|\le 1/(1-\rho(\mathcal{T}_1,\mathcal{T}_2))} \|P_{\mathcal{T}_2}\boldsymbol{M}\|_{2,\infty} \\
&= \frac{1}{1-\rho(\mathcal{T}_1,\mathcal{T}_2)} \max_{\boldsymbol{M}\in\mathcal{T}_1,\, \|\boldsymbol{M}\|=1} \|P_{\mathcal{T}_1}\boldsymbol{M} + [P_{\mathcal{T}_2} - P_{\mathcal{T}_1}]\boldsymbol{M}\|_{2,\infty} \\
&\le \frac{1}{1-\rho(\mathcal{T}_1,\mathcal{T}_2)} \left(\max_{\boldsymbol{M}\in\mathcal{T}_1,\, \|\boldsymbol{M}\|=1} \|P_{\mathcal{T}_1}\boldsymbol{M}\|_{2,\infty} + \max_{\boldsymbol{M}\in\mathcal{T}_1,\, \|\boldsymbol{M}\|=1} \|[P_{\mathcal{T}_2} - P_{\mathcal{T}_1}]\boldsymbol{M}\|_{2,\infty} \right) \\
&\le \frac{1}{1-\rho(\mathcal{T}_1,\mathcal{T}_2)} \left(\xi(\mathcal{T}_1) + \eta \max_{\|\boldsymbol{M}\|=1} \|[P_{\mathcal{T}_2} - P_{\mathcal{T}_1}]\boldsymbol{M}\| \right) \\
&= \frac{\xi(\mathcal{T}_1) + \eta\rho(\mathcal{T}_1,\mathcal{T}_2)}{1-\rho(\mathcal{T}_1,\mathcal{T}_2)},
\end{aligned}$$

where the second inequality is the triangle inequality, the third inequality follows from the definition of $\xi(\mathcal{T}_1)$ and the general bound $\|\cdot\|_{2,\infty} \le \eta\|\cdot\|_\infty \le \eta\|\cdot\|$, and the last equality uses the definition of the twisting ρ. ■

An easy corollary is the following:

Corollary C.3. *Let $\mathcal{T}, \mathcal{T}'$ be two low-rank tangent spaces (of the same dimension and to the same low-rank matrix variety) with bounded twisting $\rho(\mathcal{T}, \mathcal{T}') \le \xi(\mathcal{T})/(2\eta)$. Then, it holds that*

$$\xi(\mathcal{T}') \le 3\xi(\mathcal{T}).$$

Proof. Note that by the general bound $\xi(\mathcal{T}) \le \eta$ it follows from the assumption that $\rho(\mathcal{T}, \mathcal{T}') \le \xi(\mathcal{T})/(2\eta) \le 1/2$. Hence, the claim follows from Lemma C.2 since

$$\xi(\mathcal{T}') \le \frac{\xi(\mathcal{T}) + \eta\rho(\mathcal{T},\mathcal{T}')}{1-\rho(\mathcal{T},\mathcal{T}')} \le \frac{\xi(\mathcal{T}) + \eta\xi(\mathcal{T})/(2\eta)}{1-1/2} = 3\xi(\mathcal{T}).$$

The concludes the proof. ■

Proof of Proposition 3.2. (a) For the first claim note that

$$P_{\mathcal{J}} DH^\star(\boldsymbol{M} + \boldsymbol{N}) = (P_{\mathcal{Q}} H^\star(\boldsymbol{M} + \boldsymbol{N}), P_{\mathcal{T}'} H^\star(\boldsymbol{M} + \boldsymbol{N})).$$

We need to bound the γ-norm of this tuple. For that, we bound the respective norms for both tuple entries separately. For the first entry we calculate

$$\begin{aligned}
\|P_{\mathcal{Q}} H^\star(\boldsymbol{M}+\boldsymbol{N})\|_{\infty,2} &\geq \|P_{\mathcal{Q}} H^\star \boldsymbol{M}\|_{\infty,2} - \|P_{\mathcal{Q}} H^\star \boldsymbol{N}\|_{\infty,2} \\
&\geq \|P_{\mathcal{Q}} H^\star \boldsymbol{M}\|_{\infty,2} - \|H^\star \boldsymbol{N}\|_{\infty,2} \\
&\geq \alpha_{\mathcal{Q}} \|\boldsymbol{M}\|_{\infty,2} - \beta_{\mathcal{T}} \|\boldsymbol{N}\|_{\infty,2} \\
&\geq \alpha \|\boldsymbol{M}\|_{\infty,2} - \beta \xi(\mathcal{T}') \|\boldsymbol{N}\| \\
&\geq \alpha \|\boldsymbol{M}\|_{\infty,2} - 3\beta \xi(\mathcal{T}) \|\boldsymbol{N}\|,
\end{aligned}$$

where the first inequality is the triangle inequality, the second inequality is based on Lemma C.1, the third and fourth inequality follow from the definitions of $\alpha_{\mathcal{Q}}$, $\beta_{\mathcal{T}}$, α, β, moreover in the fourth inequality we used the definition of $\xi(\mathcal{T}')$, and the last inequality follows from Corollary C.3. Similarly, for the second entry in the tuple we calculate that

$$\begin{aligned}
\|P_{\mathcal{T}'} H^\star(\boldsymbol{M}+\boldsymbol{N})\| &\geq \|P_{\mathcal{T}'} H^\star \boldsymbol{N}\| - \|P_{\mathcal{T}'} H^\star \boldsymbol{M}\| \\
&\geq \|P_{\mathcal{T}'} H^\star \boldsymbol{N}\| - 2\|H^\star \boldsymbol{M}\| \\
&\geq \alpha_{\mathcal{T},\xi(\mathcal{T})/(2\eta)} \|\boldsymbol{N}\| - 2\beta_{\mathcal{Q}} \|\boldsymbol{M}\| \\
&\geq \alpha \|\boldsymbol{N}\| - 2\beta \|\boldsymbol{M}\| \\
&\geq \alpha \|\boldsymbol{N}\| - 2\beta \mu(\mathcal{Q}) \|\boldsymbol{M}\|_{\infty,2},
\end{aligned}$$

using the definitions of $\alpha_{\mathcal{T},\xi(\mathcal{T})/(2\eta)}$, $\beta_{\mathcal{Q}}$, α, β, and $\mu(\mathcal{Q})$. Now, in conjunction both bounds yield

$$\begin{aligned}
&\|P_{\mathcal{J}} D H^\star(\boldsymbol{M}+\boldsymbol{N})\|_\gamma \geq \max\left\{\frac{\alpha\|\boldsymbol{M}\|_{\infty,2} - 3\beta\xi(\mathcal{T})\|\boldsymbol{N}\|}{\gamma}, \alpha\|\boldsymbol{N}\| - 2\beta\mu(\mathcal{Q})\|\boldsymbol{M}\|_{\infty,2}\right\} \\
&\quad \geq \alpha \left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma - \beta \max\left\{\frac{3\xi(\mathcal{T})\|\boldsymbol{N}\|}{\gamma}, 2\mu(\mathcal{Q})\|\boldsymbol{M}\|_{\infty,2}\right\} \\
&\quad \geq \alpha \left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma - \beta \max\left\{\frac{3\xi(\mathcal{T})}{\gamma}, 2\mu(\mathcal{Q})\gamma\right\} \left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma \\
&\quad \geq \alpha \left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma - \beta \max\left\{\frac{3\xi(\mathcal{T})}{\gamma_{\min}}, 2\mu(\mathcal{Q})\gamma_{\max}\right\} \left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma \\
&\quad = \alpha \left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma - \frac{\nu\alpha}{2-\nu} \left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma = \left(\alpha - \frac{\nu\alpha}{2-\nu}\right) \left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma \qquad \text{(C.7)} \\
&\quad \geq \frac{\alpha}{2} \left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma,
\end{aligned}$$

where the second-to-last inequality follows from on $\gamma \in [\gamma_{\min}, \gamma_{\max}]$, and the final inequality is implied by $\nu \leq 1/2$.

(b) For the second claim we write

$$P_{\mathcal{J}^\perp} D H^\star(\boldsymbol{M}+\boldsymbol{N}) = (P_{\mathcal{Q}^\perp} H^\star(\boldsymbol{M}+\boldsymbol{N}), P_{\mathcal{T}'^\perp} H^\star(\boldsymbol{M}+\boldsymbol{N})).$$

Again, we bound the respective norms of both tuple entries. For the first entry we calculate

$$\begin{aligned}\|P_{\mathcal{Q}^\perp} H^\star(\boldsymbol{M}+\boldsymbol{N})\|_{\infty,2} &\leq \|P_{\mathcal{Q}^\perp} H^\star \boldsymbol{M}\|_{\infty,2} + \|P_{\mathcal{Q}^\perp} H^\star \boldsymbol{N}\|_{\infty,2}\\ &\leq \|P_{\mathcal{Q}^\perp} H^\star \boldsymbol{M}\|_{\infty,2} + \|H^\star \boldsymbol{N}\|_{\infty,2}\\ &\leq \delta_{\mathcal{Q}}\|\boldsymbol{M}\|_{\infty,2} + \beta_{\mathcal{T}}\|\boldsymbol{N}\|_{\infty,2}\\ &\leq \delta\|\boldsymbol{M}\|_{\infty,2} + \beta\xi(\mathcal{T}')\|\boldsymbol{N}\|\\ &\leq \delta\|\boldsymbol{M}\|_{\infty,2} + 3\beta\xi(\mathcal{T})\|\boldsymbol{N}\|,\end{aligned}$$

where again the first inequality is the triangle inequality, the second inequality is based on Lemma C.1, the third and fourth inequality follow from the definitions of $\delta_{\mathcal{Q}}$, $\beta_{\mathcal{T}}$, δ, β, moreover in the fourth inequality we used the definition of $\xi(\mathcal{T}')$, and the last inequality follows from Corollary C.3. The analogous calculation for the second entry is

$$\begin{aligned}\|P_{\mathcal{T}'^\perp} H^\star(\boldsymbol{M}+\boldsymbol{N})\| &\leq \|P_{\mathcal{T}'^\perp} H^\star \boldsymbol{M}\| + \|P_{\mathcal{T}'^\perp} H^\star \boldsymbol{N}\|\\ &\leq \|H^\star \boldsymbol{M}\| + \|P_{\mathcal{T}'^\perp} H^\star \boldsymbol{N}\|\\ &\leq \beta_{\mathcal{Q}}\|\boldsymbol{M}\| + \delta_{\mathcal{T},\xi(\mathcal{T})/(2\eta)}\|\boldsymbol{N}\|\\ &\leq \beta\mu(\mathcal{Q})\|\boldsymbol{M}\|_{\infty,2} + \delta\|\boldsymbol{N}\|\\ &\leq 2\beta\mu(\mathcal{Q})\|\boldsymbol{M}\|_{\infty,2} + \delta\|\boldsymbol{N}\|,\end{aligned}$$

using the definitions of $\delta_{\mathcal{T},\xi(\mathcal{T})/(2\eta)}$, $\beta_{\mathcal{Q}}$, δ, β, and $\mu(\mathcal{Q})$. Putting the two bounds together implies

$$\begin{aligned}\|P_{\mathcal{J}^\perp} DH^\star(\boldsymbol{M}+\boldsymbol{N})\|_\gamma &\leq \max\left\{\frac{\delta\|\boldsymbol{M}\|_{\infty,2} + 3\beta\xi(\mathcal{T})\|\boldsymbol{N}\|}{\gamma}, \delta\|\boldsymbol{N}\| + 2\beta\mu(\mathcal{Q})\|\boldsymbol{M}\|_{\infty,2}\right\}\\ &\leq \delta\left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma + \beta\max\left\{\frac{3\xi(\mathcal{T})\|\boldsymbol{N}\|}{\gamma}, 2\mu(\mathcal{Q})\|\boldsymbol{M}\|_{\infty,2}\right\}\\ &\leq \delta\left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma + \beta\max\left\{\frac{3\xi(\mathcal{T})}{\gamma}, 2\mu(\mathcal{Q})\gamma\right\}\left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma\\ &\leq \delta\left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma + \beta\max\left\{\frac{3\xi(\mathcal{T})}{\gamma_{\min}}, 2\mu(\mathcal{Q})\gamma_{\max}\right\}\left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma\\ &= \left(\delta + \frac{\nu\alpha}{2-\nu}\right)\left\|(\boldsymbol{M},\boldsymbol{N})\right\|_\gamma\\ &\leq \left(\delta + \frac{\nu\alpha}{2-\nu}\right)\left(\alpha - \frac{\nu\alpha}{2-\nu}\right)^{-1}\left\|P_{\mathcal{J}} DH^\star(\boldsymbol{M}+\boldsymbol{N})\right\|_\gamma\\ &= (1-\nu)\left\|P_{\mathcal{J}} DH^\star(\boldsymbol{M}+\boldsymbol{N})\right\|_\gamma,\end{aligned}$$

where once again we applied the bounds on γ in the fourth inequality, the fifth inequality uses the inequality from (C.7), and the last equality follows from the stability assumption, specifically $\nu = (1-\delta/\alpha)/2$, which implies that $\delta = (1-2\nu)\alpha$

such that

$$\begin{aligned}\delta + \frac{\nu\alpha}{2-\nu} &= (1-2\nu)\alpha + \frac{\nu\alpha}{2-\nu} \\ &= (1-\nu)\alpha + \frac{\nu\alpha - (2-\nu)\nu\alpha}{2-\nu} = (1-\nu)\left(\alpha - \frac{\nu\alpha}{2-\nu}\right).\end{aligned}$$

This finishes the proof. ■

From now on, we generally assume that the stability assumption and the γ-feasibility assumption are satisfied (and hence coupled stability holds for nearby tangent spaces).

C.3.4 Step 1: Constraining the problem to consistency

In this section, we consider the restricted problem

$$\begin{aligned}(\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}}) = \operatorname*{arg\,min}_{\boldsymbol{S},\boldsymbol{L}} \quad & \ell(\boldsymbol{S}+\boldsymbol{L}) + \lambda\left(\gamma\|\boldsymbol{S}\|_{1,2} + \|\boldsymbol{L}\|_*\right) \\ \text{s. t.} \quad & (\boldsymbol{S},\boldsymbol{L}) \in \mathcal{M}, \quad \Lambda[\boldsymbol{S}+\boldsymbol{L}] \succ \boldsymbol{0},\end{aligned} \tag{C.8}$$

where parametric and algebraic consistency of the solution are explicitly enforced by virtue of the non-convex constraint set

$$\begin{aligned}\mathcal{M} =& \Big\{(\boldsymbol{S},\boldsymbol{L}) : \boldsymbol{S} \in \mathcal{Q}(\boldsymbol{S}^\star), \operatorname{rank}(\boldsymbol{L}) \le \operatorname{rank}(\boldsymbol{L}^\star), \\ & \quad \|DH^\star(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\|_\gamma \le 9\lambda, \text{ and } \|P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L)\| \le \frac{\xi(\mathcal{T})\lambda}{\chi\|H^\star\|}\Big\}.\end{aligned}$$

Here, $\boldsymbol{\Delta}_S = \boldsymbol{S} - \boldsymbol{S}^\star$ and $\boldsymbol{\Delta}_L = \boldsymbol{L} - \boldsymbol{L}^\star$ are the errors, and the constants χ and $\|H^\star\|$ are defined in Section C.3.2. Whereas the first two constraints in $\mathcal{M}$ restrict $\boldsymbol{S}$ and $\boldsymbol{L}$ to be within the correct algebraic varieties, the last two help enforcing parametric consistency. Later in the proof, it will turn out that the additional constraints are actually inactive at the optimal solution $(\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}})$. Note that at this point we do not actually know that this solution is unique. We will show uniqueness later.

Parametric and algebraic consistency

Let us first discuss how the last two constraints in the description of $\mathcal{M}$ enforce parametric consistency. For that, we need the following lemma that shows that χ is closely related to a norm compatibility constant between the γ- and the spectral norm.

Lemma C.4. *For $\gamma \in [\gamma_{\min}, \gamma_{\max}]$ and $\boldsymbol{M} \in \operatorname{Sym}(w)$ it holds that*

$$\|D\boldsymbol{M}\|_\gamma \le \frac{\chi}{\xi(\mathcal{T})}\|\boldsymbol{M}\|,$$

where $\chi = \eta \max\{(\nu\alpha)/(3\beta(2-\nu)), 1\}$ *as we defined in Section C.3.2.*

Proof. By our choice of γ it holds

$$\begin{aligned}
\|D\boldsymbol{M}\|_\gamma &= \max\left\{\frac{\|\boldsymbol{M}\|_{\infty,2}}{\gamma}, \|\boldsymbol{M}\|\right\} \\
&\le \max\left\{\frac{\eta}{\gamma}, 1\right\} \|\boldsymbol{M}\| \\
&\le \max\left\{\frac{\eta}{\gamma_{\min}}, \frac{\eta}{\xi(\mathcal{T})}\right\} \|\boldsymbol{M}\| \\
&= \eta \max\left\{\frac{\nu\alpha}{3\beta(2-\nu)\xi(\mathcal{T})}, \frac{1}{\xi(\mathcal{T})}\right\} \|\boldsymbol{M}\| \\
&= \frac{\chi}{\xi(\mathcal{T})}\|\boldsymbol{M}\|,
\end{aligned}$$

where we used $\|\boldsymbol{M}\|_{\infty,2} \le \eta\|\boldsymbol{M}\|_\infty \le \eta\|\boldsymbol{M}\|$ in the first inequality, which also implies $\xi(\mathcal{T}) \le \eta$ that we used in the last inequality. ■

Proposition C.5 (parametric consistency). *Let* $\gamma \in [\gamma_{\min}, \gamma_{\max}]$ *and let* $(\boldsymbol{S}, \boldsymbol{L}) \in \mathcal{M}$*. Then, it holds that*

$$\|(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma \le \left(\frac{40}{\alpha} + \frac{1}{\|H^\star\|}\right)\lambda = c_2\lambda.$$

Proof. Let $\mathcal{J} = \mathcal{Q} \times \mathcal{T}$.

$$\begin{aligned}
(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L) &= P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L) + P_{\mathcal{J}^\perp}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L) \\
&= P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L) + (P_{\mathcal{Q}^\perp}(\boldsymbol{\Delta}_S), P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L)) = P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L) + (\boldsymbol{0}, P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L))
\end{aligned}$$

since $\boldsymbol{\Delta}_S \in \mathcal{Q}$. Hence, it holds by the triangle inequality that

$$\|(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma \le \|P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma + \|(\boldsymbol{0}, P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L))\|_\gamma \le \|P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma + \|P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L)\|,$$

First, the component in the orthogonal direction can be bounded as

$$\|P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L)\| \le \frac{\xi(\mathcal{T})\lambda}{\chi\|H^\star\|} \le \frac{\lambda}{\|H^\star\|}$$

which uses the fourth constraint in the definition of $\mathcal{M}$ and $\xi(\mathcal{T}) \le \chi$, which holds because $\xi(\mathcal{T}) \le \eta$ and that by definition $\eta \le \chi$. Next, the component in the tangential direction can be bounded via

$$\begin{aligned}
\|P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L))\|_\gamma &= \big\|\big(P_{\mathcal{Q}}(\boldsymbol{\Delta}_S), P_{\mathcal{T}}(\boldsymbol{\Delta}_L)\big)\big\|_\gamma \\
&\le \frac{2}{\alpha}\big\|P_{\mathcal{J}}DH^\star\big(P_{\mathcal{Q}}(\boldsymbol{\Delta}_S) + P_{\mathcal{T}}(\boldsymbol{\Delta}_L)\big)\big\|_\gamma \\
&\le \frac{4}{\alpha}\big\|DH^\star\big(P_{\mathcal{Q}}(\boldsymbol{\Delta}_S) + P_{\mathcal{T}}(\boldsymbol{\Delta}_L)\big)\big\|_\gamma
\end{aligned}$$

$$
\begin{aligned}
&= \frac{4}{\alpha}\left\|DH^\star\big(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L - P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L)\big)\right\|_\gamma \\
&\le \frac{4}{\alpha}\Big(\left\|DH^\star(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\right\|_\gamma + \left\|DH^\star P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L)\right\|_\gamma\Big) \\
&\le \frac{4}{\alpha}\Big(9\lambda + \left\|DH^\star P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L)\right\|_\gamma\Big) \\
&\le \frac{4}{\alpha}(9\lambda + \lambda) \;=\; \frac{40\lambda}{\alpha},
\end{aligned}
$$

where the first inequality is implied by Proposition 3.2 (i), the second inequality follows from Lemma C.1, the third inequality is the triangle inequality, the fourth inequality is the third constraint in the definition of $\mathcal{M}$, and the last inequality follows from

$$
\left\|DH^\star P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L)\right\|_\gamma \le \frac{\chi}{\xi(\mathcal{T})}\|H^\star P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L)\| \le \frac{\chi}{\xi(\mathcal{T})}\|H^\star\|\,\|P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L)\| \le \lambda,
$$

where here the first inequality follows from Lemma C.4, and the last inequality follows from the fourth constraint in the definition of $\mathcal{M}$. Now, the claimed bound on $\|(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma$ follows by putting together the bounds of the components in the directions of $\mathcal{J}^\perp$ and $\mathcal{J}$. ■

Note that Proposition C.5 implies parametric consistency when λ is chosen such that it goes to zero as n goes to infinity. Next, for obtaining algebraic consistency we also require the gap assumption to be satisfied.

Proposition C.6 (algebraic consistency)**.** *Under the gap assumption (and the stability and γ-feasibility assumptions) we have for $(\boldsymbol{S}, \boldsymbol{L}) \in \mathcal{M}$ that*

(i) *$\boldsymbol{S}$ has the same group support as $\boldsymbol{S}^\star$. For groups that consist of only one element, it even holds* sign consistency*. This means that these individual entries have the same sign in $\boldsymbol{S}$ and $\boldsymbol{S}^\star$.*

(ii) *$\boldsymbol{L}$ has the same rank as $\boldsymbol{L}^\star$ such that $\boldsymbol{L}$ is a smooth point in $\mathcal{L}(\mathrm{rank}(\boldsymbol{L}^\star))$. Furthermore, it holds $\boldsymbol{L} \succeq \boldsymbol{0}$.*

Proof. (i): The matrix $\boldsymbol{S}$ has the same group support as $\boldsymbol{S}^\star$ since $\boldsymbol{S} \in \mathcal{Q} = \mathcal{Q}(\boldsymbol{S}^\star)$ and

$$
\begin{aligned}
\|\boldsymbol{S} - \boldsymbol{S}^\star\|_{\infty,2} = \|\boldsymbol{\Delta}_S\|_{\infty,2} &\le \gamma\,\|(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma \le \gamma c_2\lambda \\
&\le \frac{\nu\alpha}{2\beta(2-\nu)\mu(\mathcal{Q})} c_2\lambda = \frac{c_5}{\mu(\mathcal{Q})}\lambda < s_{\min},
\end{aligned}
$$

where the first inequality holds by the definition of the γ-norm, the second inequality holds by Proposition C.5, the third inequality is implied by $\gamma \le \gamma_{\max}$, and the last inequality follows from the gap assumption. Sign consistency for groups with only one entry holds because of the sharp inequality.

(ii): We have

$$\sigma_{\min} \geq \frac{c_3\eta\lambda}{\xi(\mathcal{T})^2} = \left(\frac{6(2-\nu)}{\nu} + 1\right) c_2^2\|H^\star\|\chi\frac{\eta\lambda}{\xi(\mathcal{T})^2}$$
$$\geq 19c_2^2\|H^\star\|\chi\frac{\eta\lambda}{\xi(\mathcal{T})^2} \geq 19c_2\lambda\frac{\eta\chi}{\xi(\mathcal{T})^2} \geq 19c_2\lambda \geq 19\|\boldsymbol{\Delta}_L\|,$$

where the first inequality follows from the gap assumption, the second inequality follows from $\nu \leq 1/2$, the third inequality follows from $c_2 \geq \|H^\star\|^{-1}$, the fourth inequality follows from $\xi(\mathcal{T}) \leq \eta \leq \chi$, and the last inequality follows from Proposition C.5. Now, note that $\|\boldsymbol{\Delta}_L\|$ is the largest eigenvalue of $\boldsymbol{\Delta}_L$. Hence, the rank of $\boldsymbol{L} = \boldsymbol{L}^\star + \boldsymbol{\Delta}_L$ cannot decrease and subsequently must be the same as the one of $\boldsymbol{L}^\star$. Moreover, positive semidefiniteness of $\boldsymbol{L}$ follows from the positive semidefiniteness of $\boldsymbol{L}^\star$. ■

Further properties

We can draw some further conclusions from the properties of the elements in $\mathcal{M}$. The conclusions are also based on the following lemma.

Lemma C.7 (Propositions 2.1 and 2.2 in [Chandrasekaran et al., 2012])**.** *Let $\boldsymbol{M} \in$ Sym(w) be a rank-r matrix with smallest non-zero singular value σ. Moreover, let $\boldsymbol{\Delta}$ be a perturbation to $\boldsymbol{M}$ such that $\|\boldsymbol{\Delta}\| \leq \sigma/8$ and $\boldsymbol{M}+\boldsymbol{\Delta}$ is still a rank-r matrix. Then,*

(i) *the twisting between the two tangent spaces can be controlled via*

$$\rho(\mathcal{T}(\boldsymbol{M}+\boldsymbol{\Delta}), \mathcal{T}(\boldsymbol{M})) \leq \frac{2\|\boldsymbol{\Delta}\|}{\sigma}, \text{ and}$$

(ii) *the component of the perturbation in the normal direction can be bounded by*

$$\|P_{\mathcal{T}(\boldsymbol{M})^\perp}(\boldsymbol{\Delta})\| \leq \frac{\|\boldsymbol{\Delta}\|^2}{\sigma}.$$

First, we conclude that the fourth constraint in the definition of $\mathcal{M}$ is non-binding.

Corollary C.8. *Under the stability, γ-feasibility, and gap assumptions we have that the fourth constraint in the definition of $\mathcal{M}$ is non-binding, or more precisely*

$$\|P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L)\| \leq \frac{\xi(\mathcal{T})\lambda}{19\chi\|H^\star\|}.$$

Proof. The proof of Proposition C.6 (ii) shows that $\|\boldsymbol{\Delta}_L\| \leq \sigma_{\min}/19 \leq \sigma_{\min}/8$ and that $\boldsymbol{L} = \boldsymbol{L}^\star + \boldsymbol{\Delta}_L$ has the same rank as $\boldsymbol{L}^\star$. Hence, we can use Lemma C.7 (ii) and get

$$\|P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L)\| \leq \frac{\|\boldsymbol{\Delta}_L\|^2}{\sigma_{\min}} \leq \frac{c_2^2\lambda^2}{\sigma_{\min}} \leq \frac{c_2^2\xi(\mathcal{T})^2\lambda}{c_3\eta} \leq \frac{c_2^2\xi(\mathcal{T})\lambda}{c_3} \leq \frac{\xi(\mathcal{T})\lambda}{19\chi\|H^\star\|},$$

where the second inequality follows from Proposition C.5, the third inequality follows from the gap assumption, the fourth inequality follows from $\xi(\mathcal{T}) \leq \eta$, and the last inequality follows from the definition of c_3 and $\nu \leq 1/2$. Observe that

$$\frac{\xi(\mathcal{T})\lambda}{19\chi\|H^\star\|} < \frac{\xi(\mathcal{T})\lambda}{\chi\|H^\star\|},$$

that is, we have shown a stronger bound than the fourth constraint in the definition of $\mathcal{M}$, which therefore is non-binding. ■

We collect further properties of elements in $\mathcal{M}$ that we will use later on in the following corollary. Particularly, we show that low-rank tangent spaces to elements in $\mathcal{M}$ are close to $\mathcal{T} = \mathcal{T}(\boldsymbol{L}^\star)$, that is, the true tangent space. This is important since it allows to work with Proposition 3.2.

Corollary C.9. *Under the stability, γ-feasibility, and gap assumptions we have that*

(i) $\rho(\mathcal{T}, \mathcal{T}(\boldsymbol{L})) < \xi(\mathcal{T})/(2\eta)$,

(ii) $\left\|DH^\star P_{\mathcal{T}(\boldsymbol{L})^\perp}(\boldsymbol{L}^\star)\right\|_\gamma \leq (\lambda\nu)/(6(2-\nu))$, *and*

(iii) $\|P_{\mathcal{T}(\boldsymbol{L})^\perp}(\boldsymbol{L}^\star)\| \leq \frac{16(3-\nu)\lambda}{3\alpha(2-\nu)}$.

Proof. (i): In the proof of Corollary C.8 we have seen that $\|\boldsymbol{\Delta}_L\| \leq \sigma_{\min}/8$. Therefore, we can apply Lemma C.7 (i) such that

$$\rho(\mathcal{T}, \mathcal{T}(\boldsymbol{L})) \leq \frac{2\|\boldsymbol{\Delta}_L\|}{\sigma_{\min}} \leq \frac{2c_2\lambda}{\sigma_{\min}} \leq \frac{2c_2\xi(\mathcal{T})^2}{c_3\eta} \leq \frac{2\xi(\mathcal{T})^2}{19\chi c_2\|H^\star\|\eta} \leq \frac{2\xi(\mathcal{T})}{19\eta} < \frac{\xi(\mathcal{T})}{2\eta},$$

where the second inequality follows from Proposition C.5, the third from the gap assumption, the fourth from the definition of c_3 and $\nu \leq 1/2$, and the fifth from $\xi(\mathcal{T}) \leq \eta \leq \chi$ and $c_2 \geq \|H^\star\|^{-1}$, that is, $c_2\|H^\star\| \geq 1$.

(ii): Let σ' denote the minimum non-zero singular value of $\boldsymbol{L}$. From the proof of Proposition C.6(ii) we have $\sigma_{\min} \geq 19\|\boldsymbol{\Delta}_L\|$ and thus

$$\sigma' \geq \sigma_{\min} - \|\boldsymbol{\Delta}_L\| \geq 19\|\boldsymbol{\Delta}_L\| - \|\boldsymbol{\Delta}_L\| = 18\|\boldsymbol{\Delta}_L\|.$$

This allows us to apply Lemma C.7 (ii), where we consider $\boldsymbol{L}^\star$ as a perturbation of $\boldsymbol{L}$. In doing so we get

$$\begin{aligned}\|P_{\mathcal{T}(\boldsymbol{L})^\perp}(\boldsymbol{L}^\star)\| &= \|P_{\mathcal{T}(\boldsymbol{L})^\perp}(\boldsymbol{L} - \boldsymbol{\Delta}_L)\| = \|P_{\mathcal{T}(\boldsymbol{L})^\perp}(\boldsymbol{\Delta}_L)\| \\ &\leq \frac{\|\boldsymbol{\Delta}_L\|^2}{\sigma'} \leq \frac{c_2^2\lambda^2}{\sigma'} \leq \frac{\nu\xi(\mathcal{T})\lambda}{6(2-\nu)\chi\|H^\star\|},\end{aligned}$$

where the second inequality follows from Proposition C.5, and the last inequality follows from

$$\sigma' \geq \sigma_{\min} - \|\boldsymbol{\Delta}_L\| \geq \left(\frac{c_3\eta}{\xi(\mathcal{T})^2} - c_2\right)\lambda = \left(\left(\frac{6(2-\nu)}{\nu} + 1\right)\frac{c_2^2\chi\|H^\star\|\eta}{\xi(\mathcal{T})^2} - c_2\right)\lambda$$
$$\geq \frac{6(2-\nu)}{\nu}\frac{c_2^2\chi\|H^\star\|}{\xi(\mathcal{T})}\lambda + \left(\frac{c_2^2\chi\|H^\star\|}{\xi(\mathcal{T})} - c_2\right)\lambda \geq \frac{6(2-\nu)}{\nu}\frac{c_2^2\chi\|H^\star\|}{\xi(\mathcal{T})}\lambda,$$

where the second inequality follows from the gap assumption and Proposition C.5, the equality follows from the definition of c_3, the third inequality follows from $\xi(\mathcal{T}) \leq \eta$, and the last inequality follows from $c_2 \geq \|H^\star\|^{-1}$ and $\xi(\mathcal{T}) \leq \chi$.

Hence, we have

$$\left\|DH^\star P_{\mathcal{T}(\boldsymbol{L})^\perp}(\boldsymbol{L}^\star)\right\|_\gamma \leq \frac{\chi}{\xi(\mathcal{T})}\|H^\star P_{\mathcal{T}(\boldsymbol{L})^\perp}(\boldsymbol{L}^\star)\| \leq \frac{\chi\|H^\star\|}{\xi(\mathcal{T})}\|P_{\mathcal{T}(\boldsymbol{L})^\perp}(\boldsymbol{L}^\star)\| \leq \frac{\nu\lambda}{6(2-\nu)},$$

where the first inequality follows from Lemma C.4, the second inequality follows from the definition of $\|H^\star\|$, and the last inequality follows from the upper bound on $\|P_{\mathcal{T}(\boldsymbol{L})^\perp}(\boldsymbol{L}^\star)\|$ that we have just derived above.

(iii): We have bound $\|P_{\mathcal{T}(\boldsymbol{L})^\perp}(\boldsymbol{L}^\star)\|$ in Part (ii) of the proof. Here we use an alternative lower bound on σ', namely

$$\sigma' \geq \sigma_{\min} - \|\boldsymbol{\Delta}_L\| \geq (c_4 - c_2)\lambda \geq \frac{3\alpha c_2^2(2-\nu)}{16(3-\nu)}\lambda,$$

where the second inequality follows from the gap assumption and from Proposition C.5, and the last inequality follows from the definition of c_4. Using this alternative bound on σ' produces the claim as a consequence of

$$\|P_{\mathcal{T}(\boldsymbol{L})^\perp}(\boldsymbol{L}^\star)\| \leq \frac{\|\boldsymbol{\Delta}_L\|^2}{\sigma'} \leq \frac{c_2^2\lambda^2}{\sigma'} \leq \frac{16(3-\nu)}{3\alpha(2-\nu)}\lambda,$$

where the first inequality follows from Lemma C.7 (ii) as in the proof of Part (ii) above, the second inequality follows from Proposition C.5, and the last inequality follows from the alternative lower bound on σ'. ■

The claims in this section hold for all $(\boldsymbol{S}, \boldsymbol{L}) \in \mathcal{M}$ and hence also apply to any minimizer $(\boldsymbol{S}_\mathcal{M}, \boldsymbol{L}_\mathcal{M})$ of Problem (C.8). In the following, we work with an arbitrary fixed solution $(\boldsymbol{S}_\mathcal{M}, \boldsymbol{L}_\mathcal{M})$. Later, we show that the solution is unique. In fact, we show that it is the unique solution to the original Problem (3.7).

C.3.5 Step 2: Relaxation to tangent spaces

Our goal is to successively remove all constraints from the previously analyzed Problem (C.8). As an intermediate step, we simplify the problem by passing over to a convex relaxation of the non-convex rank constraint in the definition of $\mathcal{M}$. This rank constraint is replaced by a low-rank tangent-space constraint at the solution of Problem (C.8). Specifically, we now focus on the following tangent-space constrained problem

$$\begin{aligned} (\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}}) = \underset{\boldsymbol{S}, \boldsymbol{L}}{\arg\min} \quad & \ell(\boldsymbol{S}+\boldsymbol{L}) + \lambda\left(\gamma\|\boldsymbol{S}\|_{1,2} + \|\boldsymbol{L}\|_*\right) \\ \text{s. t.} \quad & (\boldsymbol{S}, \boldsymbol{L}) \in \mathcal{J}, \quad \Lambda[\boldsymbol{S}+\boldsymbol{L}] \succ \boldsymbol{0}, \end{aligned} \tag{C.9}$$

where $\mathcal{J} = \mathcal{Q} \times \mathcal{T}(\boldsymbol{L}_{\mathcal{M}})$. Note that the other constraints from $\mathcal{M}$ have been omitted as well. In the analysis of Problem (C.9), we proceed as follows: First, we show uniqueness of the solution. Then, we introduce some tools that will make it easier to work with optimality conditions and motivate the choice of the trade-off parameter λ. Afterwards, we show parametric consistency of the solution $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$, particularly since at this point we do not know that it is in $\mathcal{M}$. Finally, we show that $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ is indeed in $\mathcal{M}$ and thus must coincide with $(\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}})$.

Uniqueness of the solution. Here, we show how transversality of the tangent spaces implies uniqueness of the solution.

Proposition C.10 (uniqueness of the solution)**.** *Under the stability and γ-feasibility assumptions, Problem* (C.9) *is feasible and has a unique solution.*

Proof. Instead of showing uniqueness of the solution to Problem (C.9), we consider the equivalent constrained form of the problem, that is,

$$\min_{\boldsymbol{S}, \boldsymbol{L}} \ \ell(\boldsymbol{S}+\boldsymbol{L}) \quad \text{subject to} \quad (\boldsymbol{S}, \boldsymbol{L}) \in \mathcal{J},\ \Lambda[\boldsymbol{S}+\boldsymbol{L}] \succ \boldsymbol{0},\ \ \|\boldsymbol{S}\|_{1,2} \le \tau_1,\ \|\boldsymbol{L}\|_* \le \tau_2,$$

where $\tau_1, \tau_2 > 0$ are constants that depend on λ and γ. We show that this problem has a unique solution. First observe that the constraint $\Lambda[\boldsymbol{S}+\boldsymbol{L}] \succ \boldsymbol{0}$ is also implicitly enforced by a logdet-barrier from the likelihood term, see Appendix C.2. The other constraints of this problem describe a non-empty convex and compact subset of $\mathrm{Sym}(w) \times \mathrm{Sym}(w)$. Hence, the existence of a solution follows from the convexity of the objective function, which is the composition of the negative log-likelihood function and the linear addition function.

Now, uniqueness follows from strict convexity of the objective function as follows. Let $(\boldsymbol{S}, \boldsymbol{L}), (\boldsymbol{S}', \boldsymbol{L}') \in \mathcal{J}$ be distinct such that at least one of $\boldsymbol{S}-\boldsymbol{S}' \in \mathcal{Q}$ and $\boldsymbol{L}-\boldsymbol{L}' \in \mathcal{T}(\boldsymbol{L}_{\mathcal{M}})$ is non-zero. For the compound matrices $\boldsymbol{\Theta} = \boldsymbol{S}+\boldsymbol{L}$ and $\boldsymbol{\Theta}' = \boldsymbol{S}'+\boldsymbol{L}'$ we have that

$$\boldsymbol{\Theta} - \boldsymbol{\Theta}' = \boldsymbol{S} + \boldsymbol{L} - (\boldsymbol{S}' + \boldsymbol{L}') = (\boldsymbol{S} - \boldsymbol{S}') + (\boldsymbol{L} - \boldsymbol{L}') \neq \boldsymbol{0}$$

since by Proposition 3.2 and the following remark the tangent spaces are transverse, that is, $\mathcal{Q} \cap \mathcal{T}(\boldsymbol{L}_{\mathcal{M}}) = \{\boldsymbol{0}\}$. Proposition 3.2 can be applied because the necessary condition $\rho(\mathcal{T}, \mathcal{T}(\boldsymbol{L}_{\mathcal{M}})) < \xi(\mathcal{T})/(2\eta)$ holds by Corollary C.9 (i) as $\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})$ is the tangent space at $\boldsymbol{L}_{\mathcal{M}}$ with $(\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}}) \in \mathcal{M}$. Now, by Taylor expansion with the mean-value form of the remainder there exists some $t \in [0,1]$ such that

$$\begin{aligned}\ell(\mathbf{\Theta}') &= \ell(\mathbf{\Theta}) + \nabla\ell(\mathbf{\Theta})^{\mathsf{T}}(\mathbf{\Theta}-\mathbf{\Theta}') + \frac{1}{2}(\mathbf{\Theta}-\mathbf{\Theta}')^{\mathsf{T}}\nabla^2\ell\left(t\mathbf{\Theta}+(1-t)\mathbf{\Theta}'\right)(\mathbf{\Theta}-\mathbf{\Theta}')\\ &> \ell(\mathbf{\Theta}) + \nabla\ell(\mathbf{\Theta})^{\mathsf{T}}(\mathbf{\Theta}-\mathbf{\Theta}'),\end{aligned}$$

where the inequality follows from $\mathbf{\Theta}-\mathbf{\Theta}' \neq \mathbf{0}$ and the positive definiteness of the Hessian at any parameter matrix $\mathbf{\Theta}$, that is, $\boldsymbol{M}\nabla^2\ell(\mathbf{\Theta})\boldsymbol{M} > \mathbf{0}$ for all $\mathbf{0} \neq \boldsymbol{M} \in \mathrm{Sym}(w)$. This inequality establishes strict convexity of the objective as a function of $(\boldsymbol{S},\boldsymbol{L}) \in \mathcal{J}$.

Finally, for showing uniqueness suppose for a contradiction that $(\boldsymbol{S},\boldsymbol{L})$, $(\boldsymbol{S}',\boldsymbol{L}') \in \mathcal{J}$ are two distinct solutions. Then, strict convexity and equality of the objective function values imply that

$$\ell\left(\frac{1}{2}(\boldsymbol{S}+\boldsymbol{L})+\frac{1}{2}(\boldsymbol{S}'+\boldsymbol{L}')\right) < \frac{1}{2}\ell(\boldsymbol{S}+\boldsymbol{L})+\frac{1}{2}\ell(\boldsymbol{S}'+\boldsymbol{L}') = \ell(\boldsymbol{S}+\boldsymbol{L}) = \ell(\boldsymbol{S}'+\boldsymbol{L}').$$

This contradicts that $(\boldsymbol{S},\boldsymbol{L})$ and $(\boldsymbol{S}',\boldsymbol{L}')$ are solutions. Hence, there can be only one unique solution. ■

Supporting results for first-order optimality conditions. In our analysis we frequently use first-order optimality conditions. In this section, we present some tools that turn out to be useful when analyzing these optimality conditions.

First, we rewrite the gradient of the negative log-likelihood by Taylor-expansion. Using $\mathbf{\Theta}^\star = \boldsymbol{S}^\star + \boldsymbol{L}^\star$, $\mathbf{\Delta}_S = \boldsymbol{S}_{\mathcal{J}} - \boldsymbol{S}^\star$, and $\mathbf{\Delta}_L = \boldsymbol{L}_{\mathcal{J}} - \boldsymbol{L}^\star$ we have

$$\begin{aligned}\nabla\ell(\boldsymbol{S}_{\mathcal{J}}+\boldsymbol{L}_{\mathcal{J}}) &= \nabla\ell(\mathbf{\Theta}^\star+\mathbf{\Delta}_S+\mathbf{\Delta}_L) = \nabla\ell(\mathbf{\Theta}^\star) + H^\star(\mathbf{\Delta}_S+\mathbf{\Delta}_L) + R(\mathbf{\Delta}_S+\mathbf{\Delta}_L)\\ &= \nabla\ell(\mathbf{\Theta}^\star) + H^\star(\mathbf{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\mathbf{\Delta}_L) - H^\star P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star\\ &\quad + R(\mathbf{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\mathbf{\Delta}_L - P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star) \qquad \text{(C.10)}\end{aligned}$$

with the remainder $R(\mathbf{\Delta}_S+\mathbf{\Delta}_L) = \nabla\ell(\boldsymbol{S}_{\mathcal{J}}+\boldsymbol{L}_{\mathcal{J}}) - \nabla\ell(\mathbf{\Theta}^\star) - H^\star(\mathbf{\Delta}_S+\mathbf{\Delta}_L)$. In the last line, we split $\mathbf{\Delta}_L$ into its tangential and normal components for reasons that will become evident later in the proof. The remainder can be bounded using the following lemma. Remember that $r_0 > 0$ is chosen such that all parameter matrices $\mathbf{\Theta}$ in the spectral-norm ball of radius r_0 around $\mathbf{\Theta}^\star = \boldsymbol{S}^\star + \boldsymbol{L}^\star$ are feasible for the likelihood.

Lemma C.11. *Let $\mathbf{\Delta}_S \in \mathcal{Q}$ and $\gamma \in [\gamma_{\min}, \gamma_{\max}]$ such that*

$$\|(\mathbf{\Delta}_S,\mathbf{\Delta}_L)\|_\gamma \leq \max\left\{1, \frac{\nu\alpha}{2\beta(2-\nu)}\right\}^{-1}\frac{r_0}{2} = c_1.$$

Then, there exists a constant $c_0 > 0$ such that the remainder can be bounded via

$$\|DR(\mathbf{\Delta}_S+\mathbf{\Delta}_L)\|_\gamma \leq \frac{c_0}{\xi(\mathcal{T})}\|(\mathbf{\Delta}_S,\mathbf{\Delta}_L)\|_\gamma^2.$$

Proof. First note that the gradients of the negative log-likelihood and of the log-partition (normalizing) function $a(\cdot)$ differ only by a constant since $\nabla \ell(\cdot) = \nabla a(\cdot) - \hat{\boldsymbol{\Sigma}}$. In particular, it holds $H^\star = \nabla^2 \ell(\boldsymbol{\Theta}^\star) = \nabla^2 a(\boldsymbol{\Theta}^\star)$. Hence, the gradient of the log-partition function can be expanded similarly as

$$\nabla a(\boldsymbol{\Theta}^\star + \boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L) = \nabla a(\boldsymbol{\Theta}^\star) + \nabla^2 a(\boldsymbol{\Theta}^\star)(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L) + R(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)$$

with the same remainder. The remainder can be expressed using a definite-integral representation

$$\begin{aligned} R(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L) &= \nabla a(\boldsymbol{\Theta}^\star + \boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L) - \nabla a(\boldsymbol{\Theta}^\star) - \nabla^2 a(\boldsymbol{\Theta}^\star)(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L) \\ &= \int_0^1 \left[\nabla^2 a\left(\boldsymbol{\Theta}^\star + t(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\right) - \nabla^2 a(\boldsymbol{\Theta}^\star)\right](\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\, dt. \end{aligned}$$

For bounding the remainder, observe that the Hessian $\nabla^2 a$ is Lipschitz-continuous on any compact set since a is twice continuously differentiable. Let $l(r_0)$ denote the Lipschitz constant for $\nabla^2 a$ on the compact ball $B(\boldsymbol{\Theta}^\star) = \{\boldsymbol{\Theta} : \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^\star\| \le r_0\}$ such that for all $\boldsymbol{\Theta}, \boldsymbol{\Theta}' \in B(\boldsymbol{\Theta}^\star)$ it holds

$$\begin{aligned} \|\nabla^2 a(\boldsymbol{\Theta}) - \nabla^2 a(\boldsymbol{\Theta}')\| &= \max_{\boldsymbol{M} \in \mathrm{Sym}(w) :\, \|\boldsymbol{M}\|=1} \left\| \left[\nabla^2 a(\boldsymbol{\Theta}) - \nabla^2 a(\boldsymbol{\Theta}')\right] \boldsymbol{M} \right\| \\ &\le l(r_0)\|\boldsymbol{\Theta} - \boldsymbol{\Theta}'\|. \end{aligned}$$

Now, we establish that $\boldsymbol{\Theta}^\star + \boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L$ is contained in $B(\boldsymbol{\Theta}^\star)$. We do so by bounding

$$\begin{aligned} \|\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L\| &\le \|\boldsymbol{\Delta}_S\| + \|\boldsymbol{\Delta}_L\| \\ &\le \gamma\mu(\mathcal{Q})\frac{\|\boldsymbol{\Delta}_S\|_{\infty,2}}{\gamma} + \|\boldsymbol{\Delta}_L\| \\ &\le \max\left\{\gamma\mu(\mathcal{Q}), 1\right\}\left(\frac{\|\boldsymbol{\Delta}_S\|_{\infty,2}}{\gamma} + \|\boldsymbol{\Delta}_L\|\right) \\ &\le 2\max\left\{\gamma\mu(\mathcal{Q}), 1\right\}\|(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma \\ &\le 2\max\left\{\frac{\nu\alpha}{2\beta(2-\nu)}, 1\right\}\|(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma \le r_0, \end{aligned}$$

where in the third inequality we bounded the respective norms on $\boldsymbol{\Delta}_S$ and $\boldsymbol{\Delta}_L$ by the γ-norm, in the fourth inequality we used $\gamma \le \gamma_{\max}$, and in the last inequality we used the bound on $\|(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma$ from the assumption. Next, we bound the remainder in the spectral norm with the help of the Lipschitz constant $l(r_0)$ as follows

$$\begin{aligned} \|R(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\| &\le \int_0^1 \left\|\left[\nabla^2 a\left(\boldsymbol{\Theta}^\star + t(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\right) - \nabla^2 a(\boldsymbol{\Theta}^\star)\right](\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\right\| dt \\ &\le \int_0^1 \left\|\nabla^2 a\left(\boldsymbol{\Theta}^\star + t(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\right) - \nabla^2 a(\boldsymbol{\Theta}^\star)\right\| \|\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L\|\, dt \\ &\le \int_0^1 l(r_0)t\,\|\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L\|^2\, dt = \frac{l(r_0)}{2}\,\|\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L\|^2. \end{aligned}$$

This entails a bound on the γ-norm of the remainder in the following way

$$\begin{aligned}\|DR(\boldsymbol{\Delta}_S+\boldsymbol{\Delta}_L)\|_\gamma &\le \frac{\chi}{\xi(\mathcal{T})}\|R(\boldsymbol{\Delta}_S+\boldsymbol{\Delta}_L)\|\\ &\le \frac{l(r_0)\chi}{2\xi(\mathcal{T})}\|\boldsymbol{\Delta}_S+\boldsymbol{\Delta}_L\|^2\\ &\le \frac{2l(r_0)\chi}{\xi(\mathcal{T})}\max\left\{\frac{\nu\alpha}{2\beta(2-\nu)},1\right\}^2\|(\boldsymbol{\Delta}_S,\boldsymbol{\Delta}_L)\|_\gamma^2\\ &= \frac{c_0}{\xi(\mathcal{T})}\,\|(\boldsymbol{\Delta}_S,\boldsymbol{\Delta}_L)\|_\gamma^2,\end{aligned}$$

where the first inequality is a consequence of Lemma C.4, the second inequality is based on the bound on the remainder from above, and the third inequality makes use of the bound for $\|\boldsymbol{\Delta}_S+\boldsymbol{\Delta}_L\|$ also shown above. Finally, to conclude the result we define

$$c_0 = 2l(r_0)\chi\max\left\{\frac{\nu\alpha}{2\beta(2-\nu)},1\right\}^2.$$

This finishes the proof. ■

Next, for our work with the first-order optimality conditions some characterizations of norm subdifferentials will turn out to be useful.

Lemma C.12. *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$ and let $\|\cdot\|^*$ be its dual norm. Let $\boldsymbol{y}$ be in the subdifferential $\partial\|\boldsymbol{x}\|$ for some $\boldsymbol{x}\in\mathbb{R}^n$. Then, for the dual norm it holds that*

$$\|\boldsymbol{y}\|^* = \sup_{\boldsymbol{x}:\|\boldsymbol{x}\|=1}\boldsymbol{y}^\mathsf{T}\boldsymbol{x}\le 1.$$

Proof. Since $\boldsymbol{y}\in\partial\|\boldsymbol{x}\|$, the convexity of $\|\cdot\|$ implies for all $\boldsymbol{z}$ that

$$\|\boldsymbol{z}\|\ge\|\boldsymbol{x}\|+\boldsymbol{y}^\mathsf{T}(\boldsymbol{z}-\boldsymbol{x}) \quad\Longleftrightarrow\quad \boldsymbol{y}^\mathsf{T}\boldsymbol{x}-\|\boldsymbol{x}\|\ge\boldsymbol{y}^\mathsf{T}\boldsymbol{z}-\|\boldsymbol{z}\|.$$

Hence, we can also take the supremum over all $\boldsymbol{z}$ to obtain

$$\boldsymbol{y}^\mathsf{T}\boldsymbol{x}-\|\boldsymbol{x}\|\ge\sup_{\boldsymbol{z}}\left\{\boldsymbol{y}^\mathsf{T}\boldsymbol{z}-\|\boldsymbol{z}\|\right\}=\begin{cases}0, & \|\boldsymbol{y}\|^*\le 1\\ \infty, & \text{else}\end{cases}.$$

This follows because $\sup_{\boldsymbol{z}}\left\{\boldsymbol{y}^\mathsf{T}\boldsymbol{z}-\|\boldsymbol{z}\|\right\}$ is the *Fenchel conjugate* of the norm $\|\cdot\|$, which is given by the indicator function of the unit ball of the dual norm, see, for example, [Bach et al., 2012, Proposition 1.4]. Now, as the left hand side in the inequality above is always finite it must hold $\|\boldsymbol{y}\|^*\le 1$. ■

Next we record the subgradient characterizations for the $\ell_{\infty,2}$-norm and the nuclear norm. They are straight-forward adaptations of the characterizations in Section 2.2.2 to the embedding space $\mathrm{Sym}(w)$ with the group-structure prescribed by the graphical model part.

Lemma C.13. *The following holds:*

(i) *For* $\boldsymbol{S} \in \mathcal{S}(|\operatorname{gsupp}(\boldsymbol{S})|)$ *with tangent space* $\mathcal{Q}(\boldsymbol{S})$ *at* $\boldsymbol{S}$ *it holds* $\boldsymbol{Z} \in \partial\|\boldsymbol{S}\|_{1,2}$ *if and only if*

$$P_{\mathcal{Q}(\boldsymbol{S})}(\boldsymbol{Z}) = \operatorname{gsign}(\boldsymbol{S}) \quad \textit{and} \quad \|P_{\mathcal{Q}(\boldsymbol{S})^\perp}(\boldsymbol{Z})\|_{\infty,2} \leq 1,$$

where the group-sign function is defined as

$$[\operatorname{gsign}(\boldsymbol{S})]_{ij} = \begin{cases} \boldsymbol{S}_{ij}/|\boldsymbol{S}_{ij}|, & \boldsymbol{S}_{ij} \not\equiv \boldsymbol{0}, \\ \boldsymbol{0}, & \textit{otherwise} \end{cases}, \qquad i, j \in [d+q].$$

(ii) *For a rank-r matrix* $\boldsymbol{L} \in \mathcal{L}(r)$ *with tangent space* $\mathcal{T}(\boldsymbol{L})$ *at* $\boldsymbol{L}$ *it holds* $\boldsymbol{Z} \in \partial\|\boldsymbol{L}\|_*$ *if and only if*

$$P_{\mathcal{T}(\boldsymbol{L})}(\boldsymbol{Z}) = \boldsymbol{U}\boldsymbol{U}^\mathsf{T} \quad \textit{and} \quad \|P_{\mathcal{T}(\boldsymbol{L})^\perp}(\boldsymbol{Z})\| \leq 1,$$

where $\boldsymbol{L} = \boldsymbol{U}\boldsymbol{E}\boldsymbol{U}^\mathsf{T}$ *is a restricted eigenvalue decomposition of* $\boldsymbol{L}$ *with orthogonal matrix* $\boldsymbol{U} \in \mathbb{R}^{w \times r}$ *and diagonal matrix* $\boldsymbol{E} \in \mathbb{R}^{r \times r}$.

In the above characterizations, we call a subgradient (that is, an element from the subdifferential) *strictly* feasible if the inequality for the projection of the subgradient onto the normal space holds strictly.

Proof of Lemma C.13. (i) The subdifferential of a sum of convex functions is just the Minkowski sum of the respective subdifferentials. The $\ell_{1,2}$-norm is such a sum of convex functions, each of which maps onto the (vectorized) ℓ_2-norm of a *single* particular group. Elements in the subdifferential of such a function can only be non-zero in entries that belong to the particular group. Moreover, the possible values these entries can take are characterized by the subdifferential of the ℓ_2-norm which is given as follows. Let $\boldsymbol{x}$ have the same dimension as the group. Then, if $\boldsymbol{x}$ is non-zero, it holds $\partial\|\boldsymbol{x}\|_2 = \operatorname{gsign}(\boldsymbol{x})$, which corresponds to a group being in the group support of $\boldsymbol{S}$. If otherwise $\boldsymbol{x}$ is zero, we have $\partial\|\boldsymbol{x}\|_2 = \{\boldsymbol{y} : \|\boldsymbol{y}\|_2 \leq 1\}$, which corresponds to a group *not* being in the group support of $\boldsymbol{S}$. Hence, the form of the $\ell_{1,2}$-norm subdifferential follows by noting that $P_{\mathcal{Q}(\boldsymbol{S})}$ precisely projects onto the entries that belong to the group support and that $P_{\mathcal{Q}(\boldsymbol{S})^\perp}$ projects on the ones that are not contained in the group support.

(ii) See [Watson, 1992, page 41]. ■

Parametric consistency of the solution. Let us now establish parametric consistency for the unique solution of Problem (C.9).

Proposition C.14. *Next to the stability, γ-feasibility, and gap assumptions also assume*

$$\|D\nabla\ell(\boldsymbol{\Theta}^\star)\|_\gamma \le (\nu\lambda)/(6(2-\nu)) \quad \textit{and} \quad \lambda \le \min\{C_1, C_2\,\xi(\mathcal{T})\}.$$

Then, the errors $\boldsymbol{\Delta}_S = \boldsymbol{S}_{\mathcal{J}} - \boldsymbol{S}^\star$ and $\boldsymbol{\Delta}_L = \boldsymbol{L}_{\mathcal{J}} - \boldsymbol{L}^\star$ satisfy

$$\|(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma \le \frac{32(3-\nu)}{3\alpha(2-\nu)}\lambda \le c_1.$$

Proof. Similar to the proof of Proposition C.5, we decompose the error $(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)$ into its tangential part in $\mathcal{J} = \mathcal{Q} \times \mathcal{T}(\boldsymbol{L}_{\mathcal{M}})$ and its normal part in $\mathcal{J}^\perp = \mathcal{Q}^\perp \times \mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp$ and bound these parts separately. We have

$$(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L) = P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L) + P_{\mathcal{J}^\perp}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L).$$

The γ-norm of the second term is easily bound by Corollary C.9 (iii) since

$$\|P_{\mathcal{J}^\perp}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma = \left\|\text{-}(\mathbf{0}, P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star)\right\|_\gamma = \left\|P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star\right\| \le \frac{16(3-\nu)}{3\alpha(2-\nu)}\lambda.$$

We now aim at establishing the same bound in the γ-norm for the first term, that is, for $P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)$. For this more challenging task the optimality conditions of Problem (C.9) will be helpful. For that, let us first take a look at the Lagrangian of Problem (C.9). It reads as

$$\mathcal{L}(\boldsymbol{S}, \boldsymbol{L}, \boldsymbol{A}_{\mathcal{Q}^\perp}, \boldsymbol{A}_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}) = \ell(\boldsymbol{S}+\boldsymbol{L}) + \lambda(\gamma\|\boldsymbol{S}\|_{1,2} + \|\boldsymbol{L}\|_*) + \langle \boldsymbol{A}_{\mathcal{Q}^\perp}, \boldsymbol{S}\rangle + \langle \boldsymbol{A}_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}, \boldsymbol{L}\rangle,$$

where $\boldsymbol{A}_{\mathcal{Q}^\perp} \in \mathcal{Q}^\perp$ and $\boldsymbol{A}_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp} \in \mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp$ are the Lagrange multipliers. We leave the constraint $\Lambda[\boldsymbol{S}+\boldsymbol{L}] \succ \mathbf{0}$ implicit since it is enforced by the logdet-barrier that is part of the log-likelihood. Now, the optimality conditions of Problem (C.9) for the optimal solution $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ state that

$$\begin{aligned} \mathbf{0} &= \nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}}) + \boldsymbol{Z}_{1,2} + \boldsymbol{A}_{\mathcal{Q}^\perp}, & \boldsymbol{Z}_{1,2} &\in \lambda\gamma\partial\|\boldsymbol{S}_{\mathcal{J}}\|_{1,2},\\ \mathbf{0} &= \nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}}) + \boldsymbol{Z}_* + \boldsymbol{A}_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}, & \boldsymbol{Z}_* &\in \lambda\partial\|\boldsymbol{L}_{\mathcal{J}}\|_*. \end{aligned}$$

Projecting onto $\mathcal{J}$ and $\mathcal{J}^\perp$, a compact representation of these conditions is given by

$$P_{\mathcal{J}}D\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}}) = \text{-}P_{\mathcal{J}}(\boldsymbol{Z}_{1,2}, \boldsymbol{Z}_*)$$

and

$$P_{\mathcal{J}^\perp}D\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}}) = \text{-}P_{\mathcal{J}^\perp}(\boldsymbol{Z}_{1,2}, \boldsymbol{Z}_*) - (\boldsymbol{A}_{\mathcal{Q}^\perp}, \boldsymbol{A}_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}).$$

Since the Lagrange multipliers are undetermined, the second of these projected equations does not constitute a restriction on the optimal solution $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$. Instead, the optimal solution is fully characterized by the first equation that henceforth we refer to as the projected optimality condition (projected onto $\mathcal{J}$). It follows that the solution to the first equation is also unique. In fact, the important observation about the projected optimality condition is that it represents a condition on the projected error $P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L) = \big(\boldsymbol{\Delta}_S, P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L\big)$. This is an immediate consequence of

$$\begin{aligned}\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}}) &= \nabla\ell(\boldsymbol{S}_{\mathcal{J}} - \boldsymbol{S}^\star + \boldsymbol{S}^\star + \boldsymbol{L}_{\mathcal{J}} - P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{L}^\star + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{L}^\star)\\ &= \nabla\ell(\boldsymbol{\Delta}_S + \boldsymbol{S}^\star + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{L}^\star)\end{aligned}$$

which does only depend on $\boldsymbol{\Delta}_S$ and $P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L$ because $\boldsymbol{S}^\star$ and $P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{L}^\star$ are both constants.

We now show the desired bound $\|P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma \le \frac{16(3-\nu)}{3\alpha(2-\nu)}\lambda$ by constructing a map whose only fixed point is $P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)$ and that maps a γ-norm ball with radius $\frac{16(3-\nu)}{3\alpha(2-\nu)}\lambda$ onto itself. These prerequisites will then allow the application of Brouwer's fixed-point theorem [Brouwer, 1911], guaranteeing the existence of a fixed point within this small ball. This fixed point must be the unique one, that is, $P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)$. Hence, the projected error is contained in the γ-norm ball which, yields the desired bound.

To construct the map, we define $J : \mathcal{J} \to \mathcal{J}$, $(\boldsymbol{M}, \boldsymbol{N}) \mapsto P_{\mathcal{J}}DH^\star(\boldsymbol{M} + \boldsymbol{N})$. Note that the inverse operator J^{-1} is well-defined since by Proposition 3.2 the operator J is injective and thus also bijective on $\mathcal{J}$. Proposition 3.2 (i) can be applied because by Corollary C.9 (i) it holds $\rho(\mathcal{T}, \mathcal{T}(\boldsymbol{L}_{\mathcal{M}})) < \xi(\mathcal{T})/(2\eta)$. Now, setting $\boldsymbol{Z} = -P_{\mathcal{J}}(\boldsymbol{Z}_{1,2}, \boldsymbol{Z}_*)$, we consider the continuous map

$$\begin{aligned}F(\boldsymbol{M}, \boldsymbol{N}) &= (\boldsymbol{M}, \boldsymbol{N}) - J^{-1}\big(P_{\mathcal{J}}D\nabla(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}}) - \boldsymbol{Z}\big)\\ &= (\boldsymbol{M}, \boldsymbol{N}) - J^{-1}\big(P_{\mathcal{J}}D\big[\nabla\ell(\boldsymbol{\Theta}^\star) + H^\star(\boldsymbol{M} + \boldsymbol{N}) - H^\star P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star\\ &\qquad\qquad + R(\boldsymbol{M} + \boldsymbol{N} - P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star)\big] - \boldsymbol{Z}\big)\\ &= J^{-1}\big(P_{\mathcal{J}}D\big[-\nabla\ell(\boldsymbol{\Theta}^\star) + H^\star P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star\\ &\qquad\qquad - R(\boldsymbol{M} + \boldsymbol{N} - P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star)\big] + \boldsymbol{Z}\big),\end{aligned}$$

where in the second inequality we rewrote the gradient and in the last equality used that $J^{-1}P_{\mathcal{J}}DH^\star(\boldsymbol{M} + \boldsymbol{N}) = (\boldsymbol{M}, \boldsymbol{N})$ by definition. Observe that by construction any fixed point $(\boldsymbol{M}, \boldsymbol{N})$ of F must satisfy the projected optimality condition. Hence, as desired, the projected error $P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)$ is the *unique* fixed point. As outlined above, we now show that F maps a γ-norm ball with radius $\frac{16(3-\nu)}{3\alpha(2-\nu)}\lambda$ onto itself, which then allows the application of Brouwer's fixed-point theorem. For $(\boldsymbol{M}, \boldsymbol{N})$ with $\|(\boldsymbol{M}, \boldsymbol{N})\|_\gamma \le \frac{16(3-\nu)}{3\alpha(2-\nu)}\lambda$ this follows from

$$\begin{aligned}\|F(\boldsymbol{M}, \boldsymbol{N})\|_\gamma &\le \frac{2}{\alpha}\big\|P_{\mathcal{J}}D\big[\nabla\ell(\boldsymbol{\Theta}^\star) - H^\star P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star\\ &\qquad\qquad + R(\boldsymbol{M} + \boldsymbol{N} - P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star)\big] - \boldsymbol{Z}\big\|_\gamma\\ &\le \frac{2}{\alpha}\Big\{\|P_{\mathcal{J}}D\nabla\ell(\boldsymbol{\Theta}^\star)\|_\gamma + \big\|P_{\mathcal{J}}DH^\star P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star\big\|_\gamma\\ &\qquad\qquad + \big\|P_{\mathcal{J}}DR(\boldsymbol{M} + \boldsymbol{N} - P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star)\big\|_\gamma + \|\boldsymbol{Z}\|_\gamma\Big\}\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{4}{\alpha}\left\{\|D\nabla\ell(\boldsymbol{\Theta}^\star)\|_\gamma + \left\|DH^\star P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{L}^\star\right\|_\gamma + \lambda\right\} \\
&\qquad\qquad + \frac{4}{\alpha}\left\|DR(\boldsymbol{M}+\boldsymbol{N}-P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{L}^\star)\right\|_\gamma \\
&\leq \frac{4}{\alpha}\left(\frac{2(3-\nu)}{3(2-\nu)}\lambda + \frac{2(3-\nu)}{3(2-\nu)}\lambda\right) = \frac{16(3-\nu)}{3\alpha(2-\nu)}\lambda,
\end{aligned}
$$

where the first inequality follows from Proposition 3.2 (i) since the operator norm of J^{-1} is bounded by the reciprocal minimum gain of J, the second inequality is the triangle inequality, and the third inequality is implied by Lemma C.1 and

$$
\|\boldsymbol{Z}\|_\gamma = \|{-P_\mathcal{J}}(\boldsymbol{Z}_{1,2}, \boldsymbol{Z}_*)\|_\gamma = \max\left\{\gamma^{-1}\|P_\mathcal{Q}(\boldsymbol{Z}_{1,2})\|_{\infty,2}, \|P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})}(\boldsymbol{Z}_*)\|\right\} \leq 2\lambda,
$$

which holds since by the subgradient characterizations in Lemma C.13 we have that $\|P_\mathcal{Q}(\boldsymbol{Z}_{1,2})\|_{\infty,2} = \|\lambda\gamma\,\mathrm{gsign}(\boldsymbol{Z}_{1,2})\|_{\infty,2} \leq \lambda\gamma$ and since by Lemma C.12 it holds $\|\boldsymbol{Z}_*\| \leq \lambda$, which yields $\|P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})}(\boldsymbol{Z}_*)\| \leq 2\|\boldsymbol{Z}_*\| \leq 2\lambda$ in conjunction with the projection Lemma C.1. Note that for bounding $\|P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})}(\boldsymbol{Z}_*)\|$, the subgradient characterization in Lemma C.13 is not sufficient and we need Lemma C.1 because $\boldsymbol{Z}_*$ is a subgradient at $\boldsymbol{L}_\mathcal{J}$, and the tangent space at $\boldsymbol{L}_\mathcal{J}$ may be different from $\mathcal{T}(\boldsymbol{L}_\mathcal{M})$ despite having $\boldsymbol{L}_\mathcal{J} \in \mathcal{T}(\boldsymbol{L}_\mathcal{M})$. The fourth and last inequality above follows on the one hand from

$$
\|D\nabla\ell(\boldsymbol{\Theta}^\star)\|_\gamma + \left\|DH^\star P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{L}^\star\right\|_\gamma + \lambda \leq \left(2\cdot\frac{\nu}{6(2-\nu)} + 1\right)\lambda = \frac{2(3-\nu)}{3(2-\nu)}\lambda, \tag{C.11}
$$

which is a consequence of the assumed bound $\|D\nabla\ell(\boldsymbol{\Theta}^\star)\|_\gamma \leq \frac{\nu\lambda}{6(2-\nu)}$ on the gradient and $\left\|DH^\star P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{L}^\star\right\|_\gamma \leq \frac{\nu\lambda}{6(2-\nu)}$ by Corollary C.9 (ii), and on the other hand from a bound on the remainder based on Lemma C.11 given by

$$
\begin{aligned}
\left\|DR(\boldsymbol{M}+\boldsymbol{N}-P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{L}^\star)\right\|_\gamma &\leq \frac{c_0}{\xi(\mathcal{T})}\left\|(\boldsymbol{M},\boldsymbol{N}) - (\boldsymbol{0}, P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{L}^\star)\right\|_\gamma^2 \\
&\leq \frac{c_0}{\xi(\mathcal{T})}\left(\|(\boldsymbol{M},\boldsymbol{N})\|_\gamma + \|P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{L}^\star\|\right)^2 \\
&\leq \frac{c_0}{\xi(\mathcal{T})}\left(\frac{32(3-\nu)}{3\alpha(2-\nu)}\right)^2\lambda^2 \\
&\leq \frac{c_0}{\xi(\mathcal{T})}\left(\frac{32(3-\nu)}{3\alpha(2-\nu)}\right)^2\lambda\frac{3\alpha^2\nu(2-\nu)}{2^{11}c_0(3-\nu)^2}\xi(\mathcal{T}) \\
&= \frac{\nu}{6(2-\nu)}\lambda \leq \frac{2(3-\nu)}{3(2-\nu)}\lambda,
\end{aligned} \tag{C.12}
$$

where the second inequality is the triangle inequality, the third inequality uses that $(\boldsymbol{M},\boldsymbol{N})$ belongs to the γ-norm ball with radius $\frac{16(3-\nu)}{3\alpha(2-\nu)}\lambda$ and that from Corollary C.9 (iii) we have $\|P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{L}^\star\| \leq \frac{16(3-\nu)}{3\alpha(2-\nu)}\lambda$, the fourth inequality applies $\lambda \leq C_2\,\xi(\mathcal{T})$, and the last inequality uses $\nu < 4(3-\nu)$ as $\nu \leq 1/2$. Lemma C.11 can be applied since $\lambda \le C_1$ implies that

$$\left\|(\boldsymbol{M}, \boldsymbol{N} - P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star)\right\|_\gamma \le \|(\boldsymbol{M}, \boldsymbol{N})\|_\gamma + \|P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star\| \le \frac{32(3-\nu)}{3\alpha(2-\nu)}\lambda \le c_1.$$

Hence, the unique fixed point $P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)$ of F, which must be contained in the ball that F maps onto itself, indeed satisfies that $\|P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma \le \frac{16(3-\nu)}{3\alpha(2-\nu)}$. To wrap it all up, we now have

$$\|(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma \le \|P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma + \|P_{\mathcal{J}^\perp}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma \le 2 \cdot \frac{16(3-\nu)}{3\alpha(2-\nu)}\lambda \le c_1.$$

This finishes the proof. ■

Coinciding solutions. Here, we show that the solutions of the linearized Problem (C.9) and of the variety-constrained Problem (C.8) indeed coincide. Since in particular we need to show that $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}}) \in \mathcal{M}$, we begin by showing that $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ satisfies the third constraint in the description of $\mathcal{M}$.

Proposition C.15. *Under the previous assumptions, the third constraint of $\mathcal{M}$ is strictly satisfied by the solution $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ to Problem* (C.9)*, that is, for the errors $\boldsymbol{\Delta}_S = \boldsymbol{S}_{\mathcal{J}} - \boldsymbol{S}^\star$ and $\boldsymbol{\Delta}_L = \boldsymbol{L}_{\mathcal{J}} - \boldsymbol{L}^\star$ it holds that*

$$\|DH^\star(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\|_\gamma < 9\lambda.$$

Proof. We compute that

$$\begin{aligned}
\|DH^\star(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\|_\gamma &= \left\|DH^\star(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L - P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star)\right\|_\gamma \\
&\le \left\|P_{\mathcal{J}} DH^\star(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L)\right\|_\gamma \\
&\quad + \left\|P_{\mathcal{J}^\perp} DH^\star(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L)\right\|_\gamma + \left\|DH^\star P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star\right\|_\gamma \\
&\le \frac{40}{9}\lambda + \frac{40}{9}\lambda + \frac{\nu\lambda}{6(2-\nu)} \\
&\le \frac{80}{9}\lambda + \frac{1}{18}\lambda \\
&< 9\lambda,
\end{aligned}$$

where the equality follows from $\boldsymbol{\Delta}_L = P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L - P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star$ as in the proof of Proposition C.14, the first inequality is the triangle inequality, and the third inequality is implied by $\nu \le 1/2$. The second inequality follows from $\left\|DH^\star P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star\right\|_\gamma \le \frac{\nu\lambda}{6(2-\nu)}$ by Corollary C.9 (ii), from Proposition 3.2 (ii) that gives

$$\begin{aligned}
\left\|P_{\mathcal{J}^\perp} DH^\star(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L)\right\|_\gamma &\le (1-\nu)\left\|P_{\mathcal{J}} DH^\star(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L)\right\|_\gamma \\
&\le \left\|P_{\mathcal{J}} DH^\star(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L)\right\|_\gamma,
\end{aligned}$$

and finally from the rewritten form of the likelihood gradient (C.10) that produces

$$\begin{aligned}
&\left\|P_{\mathcal{J}} D H^{\star}(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L)\right\|_{\gamma} \\
&\quad = \Big\|P_{\mathcal{J}} D\Big[\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}}) - \nabla\ell(\boldsymbol{\Theta}^{\star}) + H^{\star} P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^{\perp}}\boldsymbol{L}^{\star} \\
&\qquad\qquad - R(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L - P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{L}^{\star})\Big]\Big\|_{\gamma} \\
&\quad = \left\|\boldsymbol{Z} - P_{\mathcal{J}} D\left[\nabla\ell(\boldsymbol{\Theta}^{\star}) - H^{\star} P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^{\perp}}\boldsymbol{L}^{\star} + R(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L - P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{L}^{\star})\right]\right\|_{\gamma} \\
&\quad \le \|\boldsymbol{Z}\|_{\gamma} + \|P_{\mathcal{J}} D\nabla\ell(\boldsymbol{\Theta}^{\star})\|_{\gamma} + \left\|P_{\mathcal{J}} D H^{\star} P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^{\perp}}\boldsymbol{L}^{\star}\right\|_{\gamma} \\
&\qquad\qquad + \left\|P_{\mathcal{J}} D R(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L - P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{L}^{\star})\right\|_{\gamma} \\
&\quad \le 2\left\|D R(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L - P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{L}^{\star}\right\|_{\gamma} \\
&\qquad\qquad + 2\Big[\|D\nabla\ell(\boldsymbol{\Theta}^{\star})\|_{\gamma} + \left\|D H^{\star} P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^{\perp}}\boldsymbol{L}^{\star}\right\|_{\gamma} + \lambda\Big] \\
&\quad \le 2\left[\frac{2(3-\nu)\lambda}{3(2-\nu)} + \frac{2(3-\nu)\lambda}{3(2-\nu)}\right] = \frac{8(3-\nu)}{3(2-\nu)}\lambda \le \frac{40}{9}\lambda,
\end{aligned}$$

where the second equality is implied by the projected optimality condition

$$\boldsymbol{Z} = -P_{\mathcal{J}}(\boldsymbol{Z}_{1,2}, \boldsymbol{Z}_{*}) = P_{\mathcal{J}} D\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}}),$$

the first inequality is the triangle inequality, the second inequality does some reordering and uses Lemma C.1 as well as $\|\boldsymbol{Z}\|_{\gamma} \le 2\lambda$ from the proof of Proposition C.14, the third inequality reuses the bounds (C.11) and (C.12) from the proof of Proposition C.14, which is possible since the projected error $P_{\mathcal{J}}(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L) = (\boldsymbol{\Delta}_S, P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L)$ is bounded by $\frac{16(3-\nu)}{3\alpha(2-\nu)}\lambda$, and the last inequality uses $0 < \nu \le 1/2$ and thus $(3-\nu)/(2-\nu) \le 5/3$. ■

We need two additional lemmas for proving that the solutions coincide. They are helpful for relaxing variety constraints into tangent-space constraints.

Lemma C.16 (linearization lemma)**.** *Let $\mathcal{V} \subset E$ be a variety and let $f : E \to \mathbb{R}$ be a convex continuous function. Assume that $\hat{\boldsymbol{x}}$ is a smooth point in $\mathcal{V}$ and that it minimizes the function f over the restricted domain $\mathcal{V}$, that is,*

$$\hat{\boldsymbol{x}} \in \operatorname*{arg\,min}_{\boldsymbol{x}\in\mathcal{V}} f(\boldsymbol{x}).$$

Then, $\hat{\boldsymbol{x}}$ is also a solution to the problem

$$\hat{\boldsymbol{x}} \in \operatorname*{arg\,min}_{\boldsymbol{x}\,\in\,\hat{\boldsymbol{x}}+T_{\hat{\boldsymbol{x}}}\mathcal{V}} f(\boldsymbol{x})$$

with linearized domain $\hat{\boldsymbol{x}} + T_{\hat{\boldsymbol{x}}}\mathcal{V}$, where $T_{\hat{\boldsymbol{x}}}\mathcal{V}$ is the tangent space at $\hat{\boldsymbol{x}}$ to the variety $\mathcal{V}$.

Proof. The tangent space at $\hat{\boldsymbol{x}}$ is given by the derivatives of differentiable curves passing through $\hat{\boldsymbol{x}}$, that is, $T_{\hat{\boldsymbol{x}}}\mathcal{V} = \{\gamma'(0) : \ \gamma : (-1,1) \to \mathcal{V},\, \gamma(0) = \hat{\boldsymbol{x}}\}$. Now, let $\mathbf{0} \neq \boldsymbol{\nu} \in T_{\hat{\boldsymbol{x}}}\mathcal{V}$ be a direction, and let γ be any associated curve with $\gamma : (-1,1) \to \mathcal{V}, \gamma(0) = \hat{\boldsymbol{x}}$, and $\gamma'(0) = \boldsymbol{\nu}$. By the definition of a derivative it holds $\boldsymbol{\nu} = \lim_{t\to 0} \left[\gamma(t) - \gamma(0)\right]/t$ and consequently it also holds $\hat{\boldsymbol{x}} + \boldsymbol{\nu} = \lim_{t\to 0} \left(\hat{\boldsymbol{x}} + \left[\gamma(t) - \gamma(0)\right]/t\right)$. Since $\boldsymbol{\nu} \neq \mathbf{0}$ we can assume w.l.o.g. that $\hat{\boldsymbol{x}} = \gamma(0) \neq \gamma(t)$ for all $t \neq 0$. Then, observe that for $0 < t < 1$ the points $\gamma(0)$, $\gamma(t)$, and $\hat{\boldsymbol{x}} + \left[\gamma(t) - \gamma(0)\right]/t = (1 - 1/t)\,\gamma(0) + \gamma(t)/t$ are collinear in that order (as $1/t > 1$).

Next, since $\hat{\boldsymbol{x}} = \gamma(0)$ minimizes the variety-constrained problem, the scalar function $f \circ \gamma : (-1,1) \to \mathbb{R}$ must be minimized at $t = 0$ implying that for any t it holds that $f(\hat{\boldsymbol{x}}) = f(\gamma(0)) \leq f(\gamma(t))$. Hence, by the convexity of f and collinearity it also holds that $f(\hat{\boldsymbol{x}}) = f(\gamma(0)) \leq f(\gamma(t)) \leq f\left(\hat{\boldsymbol{x}} + \left[\gamma(t) - \gamma(0)\right]/t\right)$. Now, from the continuity of f and after taking the limit $t \to 0$ it follows that $f(\hat{\boldsymbol{x}}) \leq f(\hat{\boldsymbol{x}} + \boldsymbol{\nu})$. The proof is completed by the arbitrariness of $\boldsymbol{\nu} \in T_{\hat{\boldsymbol{x}}}\mathcal{V}$. ■

Now, let us see what happens in the presence of another convex constraint.

Lemma C.17 (linearization with an additional convex constraint)**.** *Let $\mathcal{V} \subset E$ be a variety, let $f : E \to \mathbb{R}$ be a convex continuous function, and let $C \subset E$ be convex. Assume that $\hat{\boldsymbol{x}}$ is a smooth point in $\mathcal{V}$ and that it minimizes f over the domain $C \cap \mathcal{V}$, that is,*

$$\hat{\boldsymbol{x}} \in \operatorname*{arg\,min}_{\boldsymbol{x} \in C \cap \mathcal{V}} f(\boldsymbol{x}).$$

Suppose that $\hat{\boldsymbol{x}}$ does not minimize f over the linearized domain, that is,

$$\hat{\boldsymbol{x}} \notin \operatorname*{arg\,min}_{\boldsymbol{x} \in C \cap (\hat{\boldsymbol{x}} + T_{\hat{\boldsymbol{x}}}\mathcal{V})} f(\boldsymbol{x}).$$

Then, any minimizer of f over the linearized domain must be on the boundary of C.

Proof. Let $\mathbf{0} \neq \boldsymbol{\nu} \in T_{\hat{\boldsymbol{x}}}\mathcal{V}$ such that $\hat{\boldsymbol{x}} + \boldsymbol{\nu}$ minimizes the linearized problem. Let $\gamma : (-1,1) \to \mathcal{V}$ be a smooth curve with $\gamma(0) = \hat{\boldsymbol{x}}$, and $\gamma'(0) = \boldsymbol{\nu}$. First, assume for a contradiction that $f \circ \gamma : (-1,1) \to \mathbb{R}$ has its minimum at zero. Then, by the proof of the previous lemma it would follow that $f(\hat{\boldsymbol{x}} + \boldsymbol{\nu}) \geq f(\hat{\boldsymbol{x}})$, which contradicts the assumption that $\hat{\boldsymbol{x}}$ does not solve the linearized problem. Therefore, the value of f must decrease locally around $\hat{\boldsymbol{x}}$ along γ and we can assume w.l.o.g. that $f(\gamma(t)) < f(\gamma(0)) = f(\hat{\boldsymbol{x}})$ for all $t \in (0,1)$, that is, the curve enters the area where f decreases for positive t. Now, since $\hat{\boldsymbol{x}}$ solves the variety-constrained problem, it follows that $\gamma(t)$ for $t \in (0,1)$ cannot be feasible for this problem, that is, $\gamma(t) \notin C$ for $t \in (0,1)$.

Next, similarly to the proof of the previous lemma, for $0 < t < 1$ we consider the collinear points $\gamma(0) = \hat{\boldsymbol{x}} \in C$, $\gamma(t) \notin C$, and $\hat{\boldsymbol{x}} + \left[\gamma(t) - \gamma(0)\right]/t$. Then, the convexity of C implies that $\hat{\boldsymbol{x}} + \left[\gamma(t) - \gamma(0)\right]/t \notin C$. Thus, since $\hat{\boldsymbol{x}} + \left[\gamma(t) - \gamma(0)\right]/t \to \hat{\boldsymbol{x}} + \boldsymbol{\nu}$ as $t \to 0$, we have shown that there are points arbitrarily close to $\hat{\boldsymbol{x}} + \boldsymbol{\nu}$ that are not in C. Hence, the solution $\hat{\boldsymbol{x}} + \boldsymbol{\nu}$ cannot be in the interior of C. ■

Now, we can finally show that the solutions coincide.

Proposition C.18. *Under the assumptions made previously in Proposition C.14, the solutions of Problems* (C.9) *and* (C.8) *coincide, that is,* $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}}) = (\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}})$.

Proof. Let us suppose for a contradiction that the solutions are not the same. We want to apply Lemma C.17 to the product $\mathcal{V} = \mathcal{Q} \times \mathcal{L}(\operatorname{rank} \boldsymbol{L}^\star)$ and the convex set

$$C = \left\{ (\boldsymbol{S}, \boldsymbol{L}) : \|DH^\star(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\|_\gamma \le 9\lambda \text{ and } \Lambda[\boldsymbol{S} + \boldsymbol{L}] \succ \boldsymbol{0} \right\}.$$

We know that $(\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}})$ is a solution to the variety-constrained problem

$$\min_{(\boldsymbol{S}, \boldsymbol{L}) \in C \cap \mathcal{V}} \ell(\boldsymbol{S} + \boldsymbol{L}) + \lambda\left(\gamma\|\boldsymbol{S}\|_{1,2} + \|\boldsymbol{L}\|_*\right)$$

since the constraint $\|P_{\mathcal{T}^\perp}(\boldsymbol{\Delta}_L)\| \le \frac{\xi(\mathcal{T})\lambda}{\chi\|H^\star\|}$ in the description of $\mathcal{M}$ is non-binding by Corollary C.8 and dropping this constraint from $\mathcal{M}$ yields the overall constraint set $C \cap \mathcal{V}$. By Proposition C.6 we also know that the solution $\hat{\boldsymbol{x}} = (\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}})$ is a smooth point in the variety $\mathcal{V}$. Note that in this case the tangent space is given by

$$\hat{\boldsymbol{x}} + T_{\hat{\boldsymbol{x}}}\mathcal{V} = (\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}}) + \mathcal{Q} \times \mathcal{T}(\boldsymbol{L}_{\mathcal{M}}) = \mathcal{Q} \times \mathcal{T}(\boldsymbol{L}_{\mathcal{M}}) = \mathcal{J}.$$

Hence, the linearized problem is

$$\min_{(\boldsymbol{S}, \boldsymbol{L}) \in C \cap \mathcal{J}} \ell(\boldsymbol{S} + \boldsymbol{L}) + \lambda\left(\gamma\|\boldsymbol{S}\|_{1,2} + \|\boldsymbol{L}\|_*\right),$$

which is Problem (C.9) constrained to C, that is, with the additional constraint that $\|DH^\star(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\|_\gamma \le 9\lambda$. Nevertheless, this problem is also uniquely solved by the solution $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ to Problem (C.9) since by Proposition C.15 it holds for the errors $\boldsymbol{\Delta}_S = \boldsymbol{S}_{\mathcal{J}} - \boldsymbol{S}^\star$ and $\boldsymbol{\Delta}_L = \boldsymbol{L}_{\mathcal{J}} - \boldsymbol{L}^\star$ that $\|DH^\star(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\|_\gamma < 9\lambda$. Because we assumed for a contradiction that $(\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}})$ is different from $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$, it follows that $(\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}})$ does not solve the linearized problem. Therefore, we can apply Lemma C.17, which states that the solution $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ to the linearized problem must be on the boundary of C. However, the inequality from Proposition C.15 is strict, yielding the contradiction that $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ must be contained in the interior of C. Hence, it must hold $(\boldsymbol{S}_{\mathcal{M}}, \boldsymbol{L}_{\mathcal{M}}) = (\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$. ∎

A consequence of the fact that the solutions coincide is that the consistency properties from Proposition C.6 hold for the solution to Problem (C.9). In particular, we have $\operatorname{rank}(\boldsymbol{L}_{\mathcal{J}}) = \operatorname{rank}(\boldsymbol{L}^\star)$, and we have $\mathcal{T}(\boldsymbol{L}_{\mathcal{J}}) = \mathcal{T}(\boldsymbol{L}_{\mathcal{M}})$. Next, concerning parametric consistency we now have the bound $\|(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma \le \frac{32(3-\nu)}{3\alpha(2-\nu)}\lambda$ from Proposition C.14 and since we showed that the solution is also in $\mathcal{M}$, we also have the bound $\|(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma \le c_2\lambda = \left(40/\alpha + \|H^\star\|^{-1}\right)\lambda$ from Proposition C.5. An easy calculation demonstrates that the first bound is always better:

$$\frac{32(3-\nu)}{3\alpha(2-\nu)} \le \frac{40}{\alpha} \le \frac{40}{\alpha} + \frac{1}{\|H^\star\|}.$$

C.3.6 Step 3: Removing tangent-space constraints

In this section, we show that the tangent-space constraints are actually inactive at the solution $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ to the linearized Problem (C.9) such that $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ also solves the problem

$$\min_{\boldsymbol{S},\, \boldsymbol{L}} \ \ell(\boldsymbol{S}+\boldsymbol{L}) + \lambda\left(\gamma\|\boldsymbol{S}\|_{1,2} + \|\boldsymbol{L}\|_*\right) \quad \text{subject to} \quad \Lambda[\boldsymbol{S}+\boldsymbol{L}] \succ \boldsymbol{0} \tag{C.13}$$

without the tangent-space constraints. Moreover, since $\boldsymbol{L}_{\mathcal{J}} = \boldsymbol{L}_{\mathcal{M}} \succeq \boldsymbol{0}$ by Proposition C.6 (i), it then also automatically solves the original Problem (3.7). To show that $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ is also the unique solution we use the fact that it is a strictly dual feasible solution to Problem (C.13) in $\mathcal{J}$ in the sense of the subgradient characterizations from Lemma C.13 and the remark thereafter. This qualifies the solution as a primal-dual witness for the primal-dual witness proof technique.

Primal-dual witness condition. Here, we show that given a strictly dual feasible solution to Problem (C.13) that is contained in the linearized correct model space $\mathcal{J}$ there cannot be other solutions to Problem (C.13) that are not contained in $\mathcal{J}$.

Proposition C.19 (primal-dual witness). *Let $(\boldsymbol{S}_{\varnothing}, \boldsymbol{L}_{\varnothing})$ be a solution in $\mathcal{J} = \mathcal{Q} \times \mathcal{T}(\boldsymbol{L}_{\mathcal{M}})$ to Problem* (C.13) *with corresponding subgradients $\boldsymbol{Z}_{1,2} \in \lambda\gamma\partial\|\boldsymbol{S}_{\varnothing}\|_{1,2}$ and $\boldsymbol{Z}_* \in \lambda\partial\|\boldsymbol{L}_{\varnothing}\|_*$ such that it holds $\nabla\ell(\boldsymbol{S}_{\varnothing}+\boldsymbol{L}_{\varnothing})+\boldsymbol{Z}_{1,2} = \boldsymbol{0}$ and $\nabla\ell(\boldsymbol{S}_{\varnothing}+\boldsymbol{L}_{\varnothing})+\boldsymbol{Z}_* = \boldsymbol{0}$. Suppose that the subgradients satisfy the strict dual feasibility condition*

$$\|P_{\mathcal{J}^\perp}(\boldsymbol{Z}_{1,2}, \boldsymbol{Z}_*)\|_\gamma = \max\left\{\gamma^{-1}\|P_{\mathcal{Q}^\perp}\boldsymbol{Z}_{1,2}\|_{\infty,2}, \|P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{Z}_*\|\right\} < \lambda.$$

Then, all solutions to Problem (C.13) *must be in $\mathcal{J}$.*

Intuitively, the strict dual feasible condition implies that perturbing $(\boldsymbol{S}_{\varnothing}, \boldsymbol{L}_{\varnothing}) \in \mathcal{J}$ infinitesimally into a direction from the normal space $\mathcal{J}^\perp$ causes a sudden non-continuous change of the subgradients at the perturbed $(\boldsymbol{S}_{\varnothing}, \boldsymbol{L}_{\varnothing})$ by virtue of the subgradient characterizations from Lemma C.13. At the same time, the gradient of the negative log-likelihood at the perturbed $(\boldsymbol{S}_{\varnothing}, \boldsymbol{L}_{\varnothing})$ changes only infinitesimally since it is continuous. This ensures that the orthogonally perturbed solution cannot satisfy the optimality condition for Problem (C.13) and hence cannot be a solution. It is therefore that the term *witness* was coined for the solution $(\boldsymbol{S}_{\varnothing}, \boldsymbol{L}_{\varnothing})$ along with its strictly dual feasible subgradient.

Proof of Proposition C.19. Let $(\boldsymbol{S}_{\varnothing} + \boldsymbol{M}, \boldsymbol{L}_{\varnothing} + \boldsymbol{N})$ be another solution to Problem (C.13). Our goal is to show that $\boldsymbol{M} \in \mathcal{Q}$ and $\boldsymbol{N} \in \mathcal{T}(\boldsymbol{L}_{\mathcal{M}})$. First, it follows from the equality of the optimal objective function values that

$$\begin{aligned}
0 &= \ell(\boldsymbol{S}_\varnothing + \boldsymbol{M} + \boldsymbol{L}_\varnothing + \boldsymbol{N}) + \lambda\left(\gamma\|\boldsymbol{S}_\varnothing + \boldsymbol{M}\|_{1,2} + \|\boldsymbol{L}_\varnothing + \boldsymbol{N}\|_*\right) \\
&\qquad - \ell(\boldsymbol{S}_\varnothing + \boldsymbol{L}_\varnothing) - \lambda\left(\gamma\|\boldsymbol{S}_\varnothing\|_{1,2} + \|\boldsymbol{L}_\varnothing\|_*\right) \\
&\geq \langle \nabla\ell(\boldsymbol{S}_\varnothing + \boldsymbol{L}_\varnothing) + \boldsymbol{Q}_{1,2}, \boldsymbol{M}\rangle + \langle \nabla\ell(\boldsymbol{S}_\varnothing + \boldsymbol{L}_\varnothing) + \boldsymbol{Q}_*, \boldsymbol{N}\rangle \\
&= \langle \boldsymbol{Q}_{1,2} - \boldsymbol{Z}_{1,2}, \boldsymbol{M}\rangle + \langle \boldsymbol{Q}_* - \boldsymbol{Z}_*, \boldsymbol{N}\rangle \\
&= \langle P_{\mathcal{Q}^\perp}(\boldsymbol{Q}_{1,2} - \boldsymbol{Z}_{1,2}), \boldsymbol{M}\rangle + \left\langle P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}(\boldsymbol{Q}_* - \boldsymbol{Z}_*), \boldsymbol{N}\right\rangle \\
&= \langle P_{\mathcal{Q}^\perp}(\boldsymbol{Q}_{1,2} - \boldsymbol{Z}_{1,2}), P_{\mathcal{Q}^\perp}\boldsymbol{M}\rangle + \left\langle P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}(\boldsymbol{Q}_* - \boldsymbol{Z}_*), P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{N}\right\rangle,
\end{aligned}$$

where in the inequality we bounded the objective function value at $(\boldsymbol{S}_\varnothing{+}\boldsymbol{M}, \boldsymbol{L}_\varnothing{+}\boldsymbol{N})$ using a subgradient of the convex objective function at $(\boldsymbol{S}_\varnothing, \boldsymbol{L}_\varnothing)$. The subgradient of the objective function is composed of the gradient $\nabla\ell(\boldsymbol{S}_\varnothing + \boldsymbol{L}_\varnothing)$ of the negative log-likelihood (note that the derivatives w.r.t. to $\boldsymbol{S}$ and $\boldsymbol{L}$ coincide) and of some subgradients $\boldsymbol{Q}_{1,2} \in \lambda\gamma\partial\|\boldsymbol{S}_\varnothing\|_{1,2}$ and $\boldsymbol{Q}_* \in \lambda\partial\|\boldsymbol{L}_\varnothing\|_*$ that we can choose. We will make explicit choices later. In the further steps of the calculation above, we used the optimality condition for the solution $(\boldsymbol{S}_\varnothing, \boldsymbol{L}_\varnothing)$ in the second equality. In the third equality, we used that $\boldsymbol{Q}_{1,2}, \boldsymbol{Z}_{1,2} \in \lambda\gamma\partial\|\boldsymbol{S}_\varnothing\|_{1,2}$ and $\boldsymbol{Q}_*, \boldsymbol{Z}_* \in \lambda\partial\|\boldsymbol{L}_\varnothing\|_*$ which implies that their components in the respective tangent spaces $\mathcal{Q}$ and $\mathcal{T}(\boldsymbol{L}_\mathcal{M})$ coincide by the subgradient characterizations in Lemma C.13. Therefore, these components cancel out and only the projections onto the orthogonal complements of the tangent spaces remain (note the similarities to the proof of Proposition 2.5 in Appendix B.4.1).

We now choose suitable components of $\boldsymbol{Q}_{1,2}$ and $\boldsymbol{Q}_*$ in $\mathcal{Q}^\perp$ and $\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp$, respectively. By the subgradient characterization in Lemma C.13 our only restriction is that $\|P_{\mathcal{Q}^\perp}\boldsymbol{Q}_{1,2}\|_{\infty,2} \leq \lambda\gamma$ and $\|P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{Q}_*\| \leq \lambda$ must hold. We choose the orthogonal components $P_{\mathcal{Q}^\perp}\boldsymbol{Q}_{1,2} = \lambda\gamma\,\mathrm{gsign}\left(P_{\mathcal{Q}^\perp}\boldsymbol{M}\right)$ and $P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{Q}_* = \lambda\boldsymbol{O}\,\mathrm{sign}(\boldsymbol{E})\boldsymbol{O}^\top$, where $P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{N} = \boldsymbol{O}\boldsymbol{E}\boldsymbol{O}^\top$ is an eigenvalue decomposition of $P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{N}$ with orthogonal $\boldsymbol{O} \in \mathbb{R}^{w\times w}$ and diagonal $\boldsymbol{E} \in \mathbb{R}^{w\times w}$. It can be readily checked that with these choices indeed $\|P_{\mathcal{Q}^\perp}\boldsymbol{Q}_{1,2}\|_{\infty,2} \leq \lambda\gamma$ and $\|P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{Q}_*\| \leq \lambda$. Now, we continue the calculation from above with the specific subgradients

$$\begin{aligned}
&\langle P_{\mathcal{Q}^\perp}(\boldsymbol{Q}_{1,2} - \boldsymbol{Z}_{1,2}), P_{\mathcal{Q}^\perp}\boldsymbol{M}\rangle + \left\langle P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}(\boldsymbol{Q}_* - \boldsymbol{Z}_*), P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{N}\right\rangle \\
&\quad = \langle P_{\mathcal{Q}^\perp}\boldsymbol{Q}_{1,2}, P_{\mathcal{Q}^\perp}\boldsymbol{M}\rangle - \langle P_{\mathcal{Q}^\perp}\boldsymbol{Z}_{1,2}, P_{\mathcal{Q}^\perp}\boldsymbol{M}\rangle \\
&\qquad\quad + \left\langle P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{Q}_*, P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{N}\right\rangle - \left\langle P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{Z}_*, P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{N}\right\rangle \\
&\quad = \lambda\gamma\|P_{\mathcal{Q}^\perp}\boldsymbol{M}\|_{1,2} - \langle P_{\mathcal{Q}^\perp}\boldsymbol{Z}_{1,2}, P_{\mathcal{Q}^\perp}\boldsymbol{M}\rangle \\
&\qquad\quad + \lambda\|P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{N}\|_* - \left\langle P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{Z}_*, P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{N}\right\rangle \\
&\quad \geq \lambda\gamma\|P_{\mathcal{Q}^\perp}\boldsymbol{M}\|_{1,2} - \|P_{\mathcal{Q}^\perp}\boldsymbol{Z}_{1,2}\|_{\infty,2}\|P_{\mathcal{Q}^\perp}\boldsymbol{M}\|_{1,2} \\
&\qquad\quad + \lambda\|P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{N}\|_* - \|P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{Z}_*\|\|P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{N}\|_* \\
&\quad = \left(\lambda\gamma - \|P_{\mathcal{Q}^\perp}\boldsymbol{Z}_{1,2}\|_{\infty,2}\right)\|P_{\mathcal{Q}^\perp}\boldsymbol{M}\|_{1,2} + \left(\lambda - \|P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{Z}_*\|\right)\|P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{N}\|_*,
\end{aligned}$$

where the second equality follows from

$$\langle P_{\mathcal{Q}^\perp}\boldsymbol{Q}_{1,2}, P_{\mathcal{Q}^\perp}\boldsymbol{M}\rangle = \left\langle \lambda\gamma\,\mathrm{gsign}\left(P_{\mathcal{Q}^\perp}\boldsymbol{M}\right), P_{\mathcal{Q}^\perp}\boldsymbol{M}\right\rangle = \lambda\gamma\|P_{\mathcal{Q}^\perp}\boldsymbol{M}\|_{1,2}$$

and

$$\begin{aligned}\left\langle P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{Q}_*, P_{\mathcal{T}^\perp}\boldsymbol{N}\right\rangle &= \left\langle \lambda\boldsymbol{O}\operatorname{sign}(\boldsymbol{E})\boldsymbol{O}^\intercal, \boldsymbol{O}\boldsymbol{E}\boldsymbol{O}^\intercal\right\rangle \\ &= \lambda\operatorname{tr}\left(\left(\boldsymbol{O}\operatorname{sign}(\boldsymbol{E})\boldsymbol{O}^\intercal\right)^\intercal\boldsymbol{O}\boldsymbol{E}\boldsymbol{O}^\intercal\right) \\ &= \lambda\operatorname{tr}\left(\boldsymbol{O}\operatorname{sign}(\boldsymbol{E})\boldsymbol{O}^\intercal\boldsymbol{O}\boldsymbol{E}\boldsymbol{O}^\intercal\right) \\ &= \lambda\operatorname{tr}\left(\boldsymbol{O}\operatorname{sign}(\boldsymbol{E})\boldsymbol{E}\boldsymbol{O}^\intercal\right) \\ &= \lambda\operatorname{tr}\left(\boldsymbol{O}|\boldsymbol{E}|\boldsymbol{O}^\intercal\right) \\ &= \lambda\operatorname{tr}\left(|\boldsymbol{E}|\boldsymbol{O}^\intercal\boldsymbol{O}\right) = \lambda\operatorname{tr}(|\boldsymbol{E}|) = \lambda\|P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{N}\|_*,\end{aligned}$$

and the inequality follows from (the generalized) Hölder's inequality in Lemma B.3 for the respective dual norm pairs ($\ell_{\infty,2}$- and $\ell_{1,2}$-norm, and nuclear and spectral norm). In summary, we now have

$$0 \ge (\lambda\gamma - \|P_{\mathcal{Q}^\perp}\boldsymbol{Z}_{1,2}\|_{\infty,2})\,\|P_{\mathcal{Q}^\perp}\boldsymbol{M}\|_{1,2} + \left(\lambda - \|P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{Z}_*\|\right)\|P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{N}\|_*.$$

From the assumption that $\max\left\{\gamma^{-1}\|P_{\mathcal{Q}^\perp}\boldsymbol{Z}_{1,2}\|_{\infty,2}, \|P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{Z}_*\|\right\} < \lambda$ it follows that $\|P_{\mathcal{Q}^\perp}\boldsymbol{Z}_{1,2}\|_\infty < \lambda\gamma$ and $\|P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{Z}_*\| < \lambda$. Therefore, the inequality above can only be valid if $\|P_{\mathcal{Q}^\perp}\boldsymbol{M}\|_{1,2} = 0 = \|P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{N}\|_*$, that is, if $\boldsymbol{M} \in \mathcal{Q}$ and $\boldsymbol{N} \in \mathcal{T}(\boldsymbol{L}_{\mathcal{M}})$. This implies that $\boldsymbol{S}_\varnothing + \boldsymbol{M} \in \mathcal{Q}$ and $\boldsymbol{L}_\varnothing + \boldsymbol{N} \in \mathcal{T}(\boldsymbol{L}_{\mathcal{M}})$. In other words, the second solution $(\boldsymbol{S}_\varnothing + \boldsymbol{M}, \boldsymbol{L}_\varnothing + \boldsymbol{N})$ is also contained in $\mathcal{J} = \mathcal{Q} \times \mathcal{T}(\boldsymbol{L}_{\mathcal{M}})$. This finishes the proof. ■

Coinciding solutions. Finally, we show that the solution $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ to the tangent-space constrained Problem (C.9) is also the unique solution to the original Problem (3.7).

Proposition C.20 (coinciding solutions). *Assume that* $\lambda \le \min\{C_1, C_2\,\xi(\mathcal{T})\}$ *and assume that* $\|D\nabla\ell(\Theta^\star)\|_\gamma \le (\nu\lambda)/(6(2-\nu))$. *Then, under the stability, γ-feasibility, and gap assumptions, the solution $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ to the tangent-space constrained Problem* (C.9) *also uniquely solves Problem* (3.7).

Proof. It suffices to show that $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ uniquely solves Problem (C.13). This is because $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ is in $\mathcal{M}$ by Proposition C.18 and therefore it holds $\boldsymbol{L}_{\mathcal{J}} \succeq \boldsymbol{0}$ by Proposition C.6 (i). Hence, on the one hand we need to prove that $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ solves Problem (C.13), and on the other hand we must show that it is the unique solution.

We show that $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ solves Problem (C.13) by verifying the first-order optimality conditions. They require that $\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}}) \in -\lambda\gamma\partial\|\boldsymbol{S}_{\mathcal{J}}\|_{1,2}$ and $\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}}) \in -\lambda\partial\|\boldsymbol{L}_{\mathcal{J}}\|_*$. Hence, we need to check that $\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}})$ satisfies the norm-subdifferential characterizations in Lemma C.13 that can be written as

$$P_{\mathcal{J}}D\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}}) = -\lambda(\gamma\operatorname{gsign}(\boldsymbol{S}_{\mathcal{J}}), \boldsymbol{U}\boldsymbol{U}^\intercal) \quad \text{and} \quad \|P_{\mathcal{J}^\perp}D\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}})\|_\gamma \le \lambda,$$

where $\boldsymbol{L}_{\mathcal{J}} = \boldsymbol{U}\boldsymbol{E}\boldsymbol{U}^{\mathsf{T}}$ is an eigenvalue decomposition of $\boldsymbol{L}_{\mathcal{J}}$. The first condition that projects onto the components in $\mathcal{J}$ is equivalent to the optimality condition of Problem (C.9). In fact, it is the projected optimality condition from Problem (C.9) in terms of the subgradient characterizations. We already know that $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ is the unique solution in $\mathcal{J}$ to this projected optimality condition.

For showing that $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ solves Problem (C.13), it remains to establish the second condition $\|P_{\mathcal{J}^\perp} D\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}})\|_\gamma \leq \lambda$. Here, by showing the stronger sharp inequality we actually establish strict dual feasibility. This immediately implies that the only solution to Problem (C.13) must be $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ because then $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ can be used as a witness in the sense of Proposition C.19 which implies that all solutions to Problem (C.13) must be in $\mathcal{J}$. Thus, $(\boldsymbol{S}_{\mathcal{J}}, \boldsymbol{L}_{\mathcal{J}})$ must be the unique solution since we already know that it is the only solution in $\mathcal{J}$ to the projected optimality condition, that is, the optimality condition restricted to the components in $\mathcal{J}$.

The rest of the proof is dedicated to showing strict dual feasibility. We do so by making use of the Taylor expansion as in (C.10) and in the proof of Proposition C.14, namely

$$\begin{aligned}
&\|P_{\mathcal{J}^\perp} D\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}})\|_\gamma \\
&\quad = \left\|P_{\mathcal{J}^\perp} D[\nabla\ell(\boldsymbol{\Theta}^\star) + H^\star(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L) - H^\star P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star + R(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)]\right\|_\gamma \\
&\quad \leq \left\|P_{\mathcal{J}^\perp} DH^\star(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L)\right\|_\gamma \\
&\qquad\quad + \left\|P_{\mathcal{J}^\perp} D[\nabla\ell(\boldsymbol{\Theta}^\star) - H^\star P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star + R(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)]\right\|_\gamma \\
&\quad < \lambda,
\end{aligned}$$

where the first inequality is the triangle inequality, and the second one needs some more elaboration. To show it we start by applying Proposition 3.2 (ii), which yields

$$\begin{aligned}
&\left\|P_{\mathcal{J}^\perp} DH^\star(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L)\right\|_\gamma \leq (1-\nu)\left\|P_{\mathcal{J}} DH^\star(\boldsymbol{\Delta}_S + P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})}\boldsymbol{\Delta}_L)\right\|_\gamma \\
&\qquad = (1-\nu)\big\|P_{\mathcal{J}} D\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}}) \\
&\qquad\qquad\qquad - P_{\mathcal{J}} D\big[\nabla\ell(\boldsymbol{\Theta}^\star) - H^\star P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star + R(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\big]\big\|_\gamma \\
&\qquad \leq (1-\nu)\Big\{\|P_{\mathcal{J}} D\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}})\|_\gamma \\
&\qquad\qquad\qquad + \left\|P_{\mathcal{J}} D\left[\nabla\ell(\boldsymbol{\Theta}^\star) - H^\star P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star + R(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\right]\right\|_\gamma\Big\} \\
&\qquad \leq (1-\nu)\left\{\lambda + 2\left\|D\left[\nabla\ell(\boldsymbol{\Theta}^\star) - H^\star P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star + R(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\right]\right\|_\gamma\right\} \\
&\qquad \leq (1-\nu)\left\{\lambda + \frac{\nu\lambda}{2-\nu}\right\} = \frac{2\lambda(1-\nu)}{2-\nu} = \lambda - \frac{\nu\lambda}{2-\nu} < \lambda - \frac{\nu\lambda}{2(2-\nu)} \\
&\qquad \leq \lambda - \left\|D[\nabla\ell(\boldsymbol{\Theta}^\star) - H^\star P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star + R(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)]\right\|_\gamma \\
&\qquad \leq \lambda - \left\|P_{\mathcal{J}^\perp} D[\nabla\ell(\boldsymbol{\Theta}^\star) - H^\star P_{\mathcal{T}(\boldsymbol{L}_{\mathcal{M}})^\perp}\boldsymbol{L}^\star + R(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)]\right\|_\gamma,
\end{aligned}$$

where the equality is based on the Taylor expansion of the gradient, the second inequality is the triangle inequality, the third and last inequality use Lemma C.1

and

$$\|P_{\mathcal{J}} D\nabla\ell(\boldsymbol{S}_{\mathcal{J}} + \boldsymbol{L}_{\mathcal{J}})\|_\gamma = \|{-\lambda}(\gamma\,\mathrm{gsign}(\boldsymbol{S}_{\mathcal{J}}), \boldsymbol{U}\boldsymbol{U}^\mathsf{T})\|_\gamma \le \lambda,$$

and the fourth and second-to-last inequality follow from

$$\begin{aligned}
&\left\|D\left[\nabla\ell(\boldsymbol{\Theta}^\star) - H^\star P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{L}^\star + R(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\right]\right\|_\gamma \\
&\quad \le \|D\nabla\ell(\boldsymbol{\Theta}^\star)\|_\gamma + \left\|DH^\star P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{L}^\star\right\|_\gamma + \|DR(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\|_\gamma \\
&\quad \le 3\frac{\nu\lambda}{6(2-\nu)} = \frac{\nu\lambda}{2(2-\nu)},
\end{aligned}$$

which follows from the triangle inequality and from $\|D\nabla\ell(\boldsymbol{\Theta}^\star)\|_\gamma \le \frac{\nu\lambda}{6(2-\nu)}$ by assumption, from Corollary C.9 (ii) that yields $\left\|DH^\star P_{\mathcal{T}(\boldsymbol{L}_\mathcal{M})^\perp}\boldsymbol{L}^\star\right\|_\gamma \le \frac{\nu\lambda}{6(2-\nu)}$, and since the remainder too can be bounded

$$\begin{aligned}
\|DR(\boldsymbol{\Delta}_S + \boldsymbol{\Delta}_L)\|_\gamma &\le \frac{c_0}{\xi(\mathcal{T})}\|(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma^2 \\
&\le \frac{c_0\lambda^2}{\xi(\mathcal{T})}\left(\frac{32(3-\nu)}{3\alpha(2-\nu)}\right)^2 \\
&\le c_0\lambda\left(\frac{32(3-\nu)}{3\alpha(2-\nu)}\right)^2 C_2 = c_0\lambda\left(\frac{32(3-\nu)}{3\alpha(2-\nu)}\right)^2 \frac{3\alpha^2\nu(2-\nu)}{2^{11}c_0(3-\nu)^2} \\
&= \frac{\nu\lambda}{6(2-\nu)},
\end{aligned}$$

where the first inequality uses Lemma C.11, which is possible since from Proposition C.14 it follows that $\|(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma \le \frac{32(3-\nu)}{3\alpha(2-\nu)}\lambda \le c_1$. This also explains the second inequality. Finally, the last inequality follows from $\lambda \le C_2\,\xi(\mathcal{T})$. This finishes the proof. ■

This concludes also the proof of Theorem 3.3 since we have shown that under the assumptions from Theorem 3.3, particularly the assumed bound on the gradient, the solution to Problem (3.7) is algebraically consistent in light of Proposition C.6 and parametrically consistent in the sense that $\|(\boldsymbol{\Delta}_S, \boldsymbol{\Delta}_L)\|_\gamma \le \frac{32(3-\nu)}{3\alpha(2-\nu)}\lambda$ by Proposition C.14.

C.3.7 Proof of corollaries via probabilistic analyses

In this section, we prove Corollaries 3.4, 3.5, and 3.6. For each proof, a probabilistic analysis is necessary with the aim of bounding the sampling error in the γ-norm. This sampling error is given as the gradient

$$\nabla\ell(\boldsymbol{\Theta}^\star) = \nabla\ell(\boldsymbol{S}^\star + \boldsymbol{L}^\star) = \nabla\left(a(\boldsymbol{\Theta}) - \langle\boldsymbol{\Theta}, \hat{\boldsymbol{\Sigma}}\rangle\right)_{|\boldsymbol{\Theta}=\boldsymbol{\Theta}^\star} = \boldsymbol{\Sigma}^\star - \hat{\boldsymbol{\Sigma}}$$

of the respective likelihood functions. Here, $\hat{\boldsymbol{\Sigma}} = n^{-1}\sum_{k=1}^n(\overline{\boldsymbol{x}}^{(k)}, \boldsymbol{y}^{(k)})\left[(\overline{\boldsymbol{x}}^{(k)}, \boldsymbol{y}^{(k)})\right]^\mathsf{T}$ is the empirical and $\boldsymbol{\Sigma}^\star = \mathbb{E}\left[(\overline{\boldsymbol{x}}, \boldsymbol{y})[(\overline{\boldsymbol{x}}, \boldsymbol{y})]^\mathsf{T}\right]$ is the population version of the second-moment matrix, where the expectation is taken w.r.t. the true pairwise CG distribution with parameter matrix $\boldsymbol{\Theta}^\star = \boldsymbol{S}^\star + \boldsymbol{L}^\star$.

Gaussian case: probabilistic analysis

Proof of Corollary 3.4. Note that in this special case $w = q$. By [Chandrasekaran et al., 2012, Corollary 5.5] it holds

$$\mathbb{P}\left(\|\boldsymbol{\Sigma}^\star - \hat{\boldsymbol{\Sigma}}\| \leq \sqrt{128q\|\boldsymbol{\Sigma}^\star\|^2/n}\right) \leq 1 - 2\exp(-q),$$

provided that $n \geq 2q$. Hence, with the choice of $f(q) = 128q\|\boldsymbol{\Sigma}^\star\|^2$ and

$$n > \max\left\{C_5^2\, f(w)\, [\xi(\mathcal{T}) \min\{C_1, C_2\, \xi(\mathcal{T})\}]^{-2}, 2q\right\}$$

the lower bound (3.10) on n is satisfied, and we have with probability at least $1 - 2\exp(-q)$ that

$$\begin{aligned}\|D\nabla\ell(\boldsymbol{\Theta}^\star)\|_\gamma &\leq \frac{\chi}{\xi(\mathcal{T})}\|\nabla\ell(\boldsymbol{\Theta}^\star)\| = \frac{\chi}{\xi(\mathcal{T})}\|\boldsymbol{\Sigma}^\star - \hat{\boldsymbol{\Sigma}}\|\\ &\leq \frac{\chi}{\xi(\mathcal{T})}\sqrt{\frac{128q\|\boldsymbol{\Sigma}^\star\|^2}{n}} = \frac{\chi}{\xi(\mathcal{T})}\sqrt{\frac{f(q)}{n}} = \frac{\nu}{6(2-\nu)}\lambda_{n,f(q)},\end{aligned}$$

where the first inequality follows from Lemma C.4, and we used the definition

$$\lambda_{n,f(q)} = C_5/\xi(\mathcal{T})\,\sqrt{f(q)/n}$$

with $C_5 = 6(2-\nu)\chi/\nu$ in the last equality. An application of Theorem 3.3 with $\lambda = \lambda_{n,f(q)}$ concludes the proof of Corollary 3.4. ■

Discrete case: probabilistic analysis

For the proof, we will use the following lemma that is based on [Vershynin, 2010, Corollary 5.52]. The constant c_I that appears in the lemma is independent of d and is defined in [Vershynin, 2010, Corollary 5.52].

Lemma C.21. *Let $\boldsymbol{\Sigma}^\star$ and $\hat{\boldsymbol{\Sigma}}$ be defined as before, where we assume $\boldsymbol{\Sigma}^\star$ to be invertible. Let $\kappa \geq 1$. Then, assume that it holds $n > c_I\kappa\|\boldsymbol{\Sigma}^\star\|^{-1}d\log m$ for an absolute constant c_I. Then, if we set $\delta_n = \sqrt{c_I\kappa\|\boldsymbol{\Sigma}^\star\|d\log m/n}$ it holds that*

$$\mathbb{P}\left(\|\hat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\| > \delta_n\right) \leq m^{-\kappa}.$$

Proof. First, observe that $\|\overline{\boldsymbol{x}}\|_2^2 \leq d$ for all $\boldsymbol{x} \in \mathcal{X}$ since any concatenated indicator representation contains at most d ones. We build the remaining proof upon the classical result [Vershynin, 2010, Corollary 5.52] for *bounded* random vectors. This result implies that for any $0 < \varepsilon < 1$ we have

$$\mathbb{P}\left(\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star\| > \varepsilon\|\mathbf{\Sigma}^\star\|\right) \leq \exp\left(-\kappa + \log m\right) = m^{-\kappa}$$

as long as the number of samples satisfies $n \geq c_I \kappa \varepsilon^{-2}\|\mathbf{\Sigma}^\star\|^{-1} d \log m$ (the absolute constant c_I is independent of d and m and is defined in [Vershynin, 2010, Corollary 5.52]). We intend to use this result with

$$\varepsilon = \varepsilon_n = \frac{\delta_n}{\|\mathbf{\Sigma}^\star\|} = \sqrt{\frac{c_I \kappa d \log m}{\|\mathbf{\Sigma}^\star\| n}}.$$

For ε_n chosen in this way, one can check that the lower bound on n required by [Vershynin, 2010, Corollary 5.52] is trivially satisfied. It also follows that $\varepsilon_n < 1$ by plugging in the lower bound on n that we assumed for this lemma:

$$\varepsilon_n^2 = \frac{c_I \kappa d \log m}{\|\mathbf{\Sigma}^\star\| n} < \frac{c_I \kappa d \log m}{\|\mathbf{\Sigma}^\star\| c_I \kappa \|\mathbf{\Sigma}^\star\|^{-1} d \log m} = 1.$$

Now, applying [Vershynin, 2010, Corollary 5.52] yields the claim

$$\mathbb{P}\left(\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star\| > \varepsilon_n\|\mathbf{\Sigma}^\star\|\right) = \mathbb{P}\left(\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}^\star\| > \delta_n\right) \leq m^{-\kappa}.$$

This finishes the proof. ■

We are now ready to prove Corollary 3.5.

Proof of Corollary 3.5. Note that here, the scaling f depends on the number of discrete variables d and m. Let $f(d,m) = c_I \kappa \|\mathbf{\Sigma}^\star\| d \log m$ and

$$n > \max\left\{C_5^2\, f(d,m)\left[\xi(\mathcal{T}) \min\{C_1, C_2\, \xi(\mathcal{T})\}\right]^{-2}, c_I \kappa \|\mathbf{\Sigma}^\star\|^{-1} d \log m\right\}.$$

That way, the lower Bound (3.10) on n is satisfied (and the second term does not change the asymptotics). Now, it follows that the required bound on the gradient also holds with high probability since

$$\|D\nabla\ell(\mathbf{\Theta}^\star)\|_\gamma \leq \frac{\chi}{\xi(\mathcal{T})}\|\nabla\ell(\mathbf{\Theta}^\star)\| \leq \frac{\chi}{\xi(\mathcal{T})}\delta_n = \frac{\chi}{\xi(\mathcal{T})}\sqrt{\frac{f(d,m)}{n}} = \frac{\nu}{6(2-\nu)}\lambda_{n,f(d,m)},$$

where the first inequality is a consequence of Lemma C.4, and the second inequality holds with probability at least $1 - m^{-\kappa}$ by Lemma C.21 which can be applied because of the assumed lower bound on n (second term in the max). This concludes the proof of Corollary 3.5 after an application of Theorem 3.3 with $\lambda = \lambda_{n,f(d,m)}$. ■

Mixed case: probabilistic analysis

Bounding the spectral norm of the sampling error turns out to be quite challenging for the CG distribution that includes both discrete and unbounded continuous variables. In the following, we derive bounds similar to the ones given in [Lee et al., 2015] for the maximum norm. For that, we first present a lemma that is based on [Vershynin, 2010, Corollary 5.17].

Lemma C.22 (concentration lemma by exponential type tail and union bound)**.** *For some constant K independent of n and for some $c_M > 0$, for any $\varepsilon > 0$ the random vector $\nabla\ell(\boldsymbol{\Theta}^\star) = \boldsymbol{\Sigma}^\star - \hat{\boldsymbol{\Sigma}}$ satisfies*

$$\mathbb{P}\left(\|\nabla\ell(\boldsymbol{\Theta}^\star)\|_\infty > \varepsilon\right) \le 2\exp\left(2\log w - c_M n \min\left\{\frac{\varepsilon^2}{K^2}, \frac{\varepsilon}{K}\right\}\right).$$

Proof. Observe that the derivative w.r.t. the (i,j)-th matrix entry (this time not the group) satisfies

$$\frac{\partial\ell(\boldsymbol{\Theta}^\star)}{\partial\Theta_{ij}} = \frac{1}{n}\sum_{k=1}^{n}\left(\mathbb{E}\left[\Sigma_{ij}^\star\right] - \left((\overline{\boldsymbol{x}}^{(k)}, \boldsymbol{y}^{(k)})\left[(\overline{\boldsymbol{x}}^{(k)}, \boldsymbol{y}^{(k)})\right]^\mathsf{T}\right)_{ij}\right), \qquad i,j \in [w].$$

This is a sum of i.i.d. centered subexponential random variables by [Lee et al., 2015, Lemma B.1]. By applying the subexponential bound from [Vershynin, 2010, Corollary 5.17] we see that

$$\mathbb{P}\left(\left|\frac{\partial\ell(\boldsymbol{\Theta}^\star)}{\partial\Theta_{ij}}\right| > \varepsilon\right) \le 2\exp\left(-c_M n\min\left\{\frac{\varepsilon^2}{K_{ij}^2}, \frac{\varepsilon}{K_{ij}}\right\}\right), \qquad i,j \in [w],$$

where K_{ij} is the Orlicz 1-norm of the random variable $\mathbb{E}\left[\Sigma_{ij}^\star\right] - (\overline{\boldsymbol{x}}, \boldsymbol{y})_i(\overline{\boldsymbol{x}}, \boldsymbol{y})_j$, and c_M is an absolute constant that is defined in [Vershynin, 2010, Corollary 5.17]. Let $K = \max_{i,j} K_{ij}$. By a union bound we have

$$\begin{aligned}\mathbb{P}\left(\|\nabla\ell(\boldsymbol{\Theta}^\star)\|_\infty > \varepsilon\eta^{-1}\right) &= \mathbb{P}\left(\text{for some } (i,j) \in [w]\times[w] : \left|\frac{\partial\ell(\boldsymbol{\Theta}^\star)}{\partial\Theta_{ij}}\right| > \varepsilon\right) \\ &\le \sum_{i,j=1}^{w}\mathbb{P}\left(\left|\frac{\partial\ell(\boldsymbol{\Theta}^\star)}{\partial\Theta_{ij}}\right| > \varepsilon\right) \\ &\le 2w^2\exp\left(-c_M n\min\left\{\frac{\varepsilon^2}{K^2}, \frac{\varepsilon}{K}\right\}\right) \\ &= 2\exp\left(2\log w - c_M n\min\left\{\frac{\varepsilon^2}{K^2}, \frac{\varepsilon}{K}\right\}\right).\end{aligned}$$

This finishes the proof. ■

As a consequence we have:

Corollary C.23 (bound on the CG log-likelihood gradient)**.** *Assume that the data is drawn from the true pairwise CG distribution with interaction matrix $\boldsymbol{\Theta}^\star = \boldsymbol{S}^\star + \boldsymbol{L}^\star$.*

Let $\kappa > 0$ and $n \geq 2c_M^{-1}(1+\kappa/2)\log w$. Then, it holds that

$$\mathbb{P}\left(\|D\nabla\ell(\mathbf{\Theta}^\star)\|_\gamma > \sqrt{\frac{2c_M^{-1}K^2(1+\kappa/2)w^2\log w}{n}}\frac{\chi}{\xi(\mathcal{T})}\right) \leq 2w^{-\kappa}.$$

Proof. With some more generality we show that if it holds $n \geq K^{-2}g(w)$ for some function $g(w) > 0$, then we have that

$$\mathbb{P}\left(\|D\nabla\ell(\mathbf{\Theta}^\star)\|_\gamma > \sqrt{\frac{w^2 g(w)}{n}}\frac{\chi}{\xi(\mathcal{T})}\right) \leq 2\exp\left(2\log w - c_M K^{-2}g(w)\right).$$

To prove this, set $\varepsilon_n = \sqrt{g(w)/n}$. Then,

$$\begin{aligned}\mathbb{P}\left(\|D\nabla\ell(\mathbf{\Theta}^\star)\|_\gamma > w\varepsilon_n\frac{\chi}{\xi(\mathcal{T})}\right) &\leq \mathbb{P}\left(\|\nabla\ell(\mathbf{\Theta}^\star)\| > w\,\varepsilon_n\right) \leq \mathbb{P}\left(\|\nabla\ell(\mathbf{\Theta}^\star)\|_\infty > \varepsilon_n\right)\\ &\leq 2\exp\left(2\log w - c_M n\min\left\{\frac{\varepsilon_n^2}{K^2}, \frac{\varepsilon_n}{K}\right\}\right)\\ &= 2\exp\left(2\log w - c_M n\frac{\varepsilon_n^2}{K^2}\right)\\ &= 2\exp\left(2\log w - c_M K^{-2}g(w)\right),\end{aligned}$$

where the first inequality follows from Lemma C.4, the second inequality follows from the general bound $\|\cdot\| \leq w\|\cdot\|_\infty$, and the third inequality follows from the preceding Lemma C.22. If $\varepsilon_n = \sqrt{g(w)/n} \leq K$, that is, if $n \geq K^{-2}g(w)$, then the first term in the minimum is active: This yields the first equality. Now, the claim follows with the specific choice $g(w) = 2c_M^{-1}K^2(1+\kappa/2)\log w$ since then

$$2\log w - c_M K^{-2}g(w) = 2\log w - 2(1+\kappa/2)\log w = -\kappa\log w.$$

Moreover, the lower bound on n above becomes $n \geq K^{-2}g(w) = 2c_M^{-1}(1+\kappa/2)\log w$. This finishes the proof. ■

In the previous result, we used the weak bound $\|\cdot\| \leq w\|\cdot\|_\infty$ which caused an additional factor of w in comparison to the sampling error bound in the maximum norm for sparse graphical model estimation from [Lee and Hastie, 2015; Lee et al., 2015]. We conjecture that the spectral norm actually can be bounded with the same strength as the maximum norm. However, classical results for spectral norm bounds of the sampling error from random matrix theory, see [Vershynin, 2010] and [Adamczak et al., 2010], typically require subgaussian, log-concavity, or boundedness assumptions on the distribution. Unfortunately, given that here the observed data consists of both discrete and unbounded quantitative variables all these assumptions are not satisfied. Let us now prove Corollary 3.6.

Proof of Corollary 3.6. Let $f(w) = 2c_M^{-1}K^2(1+\kappa/2)w^2 \log w$ and

$$n > \max\left\{C_5^2\, f(w)\, [\xi(\mathcal{T}) \min\{C_1, C_2\, \xi(\mathcal{T})\}]^{-2}, 2c_M^{-1}(1+\kappa/2)\log w\right\}.$$

Again, the lower bound (3.10) is satisfied (and the asymptotics remain the same). Moreover, by Corollary C.23 (ii) it follows with probability at least $1 - 2w^{-\kappa}$ that

$$\|D\nabla\ell(\boldsymbol{\Theta}^{\star})\|_{\gamma} \leq \sqrt{\frac{f(w)}{n}}\frac{\chi}{\xi(\mathcal{T})} = \frac{\nu}{6(2-\nu)}\lambda_{n,f(w)}.$$

Hence, the bound on the gradient required by Theorem 3.3 holds with high probability and thus the proof of Corollary 3.6 can be concluded by an application of Theorem 3.3 with $\lambda = \lambda_{n,f(w)}$. ■

Bibliography

R. Adamczak, A. Litvak, A. Pajor, and N. Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2):535–561, 2010.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends in Machine Learning*, 4(1):1–106, 2012.

D. J. Bartholomew and M. Knott. *Latent variable models and factor analysis.* Arnold London, 1999.

H. P. Benson. An outer approximation algorithm for generating all efficient extreme points in the outcome set of a multiple objective linear programming problem. *Journal of Global Optimization*, 13(1):1–24, 1998.

J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 24(3):179–195, 1975.

K. Blechschmidt, J. Giesen, and S. Laue. Tracking approximate solutions of parameterized optimization problems over multi-dimensional (hyper-)parameter domains. In *International Conference on Machine Learning (ICML)*, pages 438–447, 2015.

B. Bollobás. *Random graphs.* Cambridge University Press, 2001.

S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

D. Bremner, K. Fukuda, and A. Marzetta. Primal-dual methods for vertex and facet enumeration. *Discrete & Computational Geometry*, 20(3):333–357, 1998.

L. E. J. Brouwer. Über Abbildungen von Mannigfaltigkeiten. *Mathematische Annalen*, 71(1):97–115, 1911.

E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

E. J. Candès, X. Li, Y. Ma, and J. Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

G. Casella and E. I. George. Explaining the Gibbs sampler. *The American Statistician*, 46(3):167–174, 1992.

V. Chandrasekaran, S. Sanghavi, P. A. Parrilo, and A. S. Willsky. Rank-sparsity incoherence for matrix decomposition. *SIAM Journal on Optimization*, 21(2): 572–596, 2011.

V. Chandrasekaran, P. A. Parrilo, and A. S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012.

J. Chen, J. Zhou, and J. Ye. Integrating low-rank and group-sparse structures for robust multi-task learning. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 42–50, 2011.

S. S. Chen, D. L. Donoho, and M. A. Saunders. Atomic decomposition by basis pursuit. *SIAM review*, 43(1):129–159, 2001.

Y. Chen, X. Li, J. Liu, and Z. Ying. A fused latent and graphical model for multivariate binary data. Technical report, arXiv preprint arXiv:1606.08925, 2016.

Y. Chen, X. Li, J. Liu, and Z. Ying. Robust measurement via a fused latent and graphical item response theory model. *Psychometrika*, pages 1–25, 2018.

J. Cheng, T. Li, E. Levina, and J. Zhu. High-dimensional mixed graphical models. *Journal of Computational and Graphical Statistics*, 26(2):367–378, 2017.

S. Dempe. *Foundations of Bilevel Programming*. Springer, 2002.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39 (1):1–38, 1977.

D. Dua and E. Karra Taniskidou. UCI Machine Learning Repository, 2017. URL `http://archive.ics.uci.edu/ml`.

M. Dudík, S. J. Phillips, and R. E. Schapire. Performance guarantees for regularized maximum entropy density estimation. In *Conference on Learning Theory (COLT)*, pages 472–486, 2004.

C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

S. E. Embretson and S. P. Reise. *Item Response Theory*. Psychology Press, 2013.

ESA. Sentinel-2 mission, tile t16tfl (2019-2020), accessed: 2020-08-06., 2020. URL `https://scihub.copernicus.eu/`.

S. M. Fazel. *Matrix rank minimization with applications*. PhD thesis, Standford University, 2003.

R. Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

C. Fox. *An introduction to the calculus of variations*. Courier Corporation, 1987.

J.-M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders. The Amsterdam library of object images. *International Journal of Computer Vision*, 61(1):103–112, 2005.

J. Giesen, S. Laue, A. Löhne, and C. Schneider. Using Benson's algorithm for regularization parameter tracking. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3689–3696, 2019a.

J. Giesen, F. Nussbaum, and C. Schneider. Efficient regularization parameter selection for latent variable graphical models via bi-level optimization. In *IJCAI*, pages 2378–2384, 2019b.

N. Groll and R. Weisse. coastdat-2 North Sea wave hindcast for the period 1949-2014 performed with the wave model wam. *World Date Center for Climate (WDCC) at DKRZ*, 2016.

M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, and B. Reed. *Probabilistic methods for algorithmic discrete mathematics*, volume 16. Springer Science & Business Media, 2013.

N. Halko, P.-G. Martinsson, and J. A. Tropp. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Review*, 53(2):217–288, 2011.

P. Hansen, B. Jaumard, and G. Savard. New branch-and-bound rules for linear bilevel programming. *SIAM Journal on Scientific and Statistical Computing*, 13 (5):1194–1217, 1992.

R. M. Haralick. Statistical and structural approaches to texture. *Proceedings of the IEEE*, 67(5):786–804, 1979.

J. Harris. *Algebraic geometry: a first course*, volume 133. Springer Science & Business Media, 2013.

Helmholtz Centre for Materials and Coastal Research. coastdat-1 waves north sea wave spectra hindcast (1948-2007), 2012. Geesthacht.

C. Higuera, K. J. Gardiner, and K. J. Cios. Self-organizing feature maps identify proteins critical to learning in a mouse model of down syndrome. *PloS ONE*, 10 (6):1–28, 2015.

R. A. Horn and C. R. Johnson. *Matrix analysis*. Cambridge university press, 2012.

H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, 24(6):417–441, 1933.

B. Huang, C. Mu, D. Goldfarb, and J. Wright. Provable low-rank tensor recovery. *Optimization-Online*, 4252(2):455–500, 2014.

D. Inouye, P. Ravikumar, and I. Dhillon. Square root graphical models: Multivariate generalizations of univariate exponential families that permit positive dependencies. In *International Conference on Machine Learning*, pages 2445–2453. PMLR, 2016.

E. Ising. Contribution to the theory of ferromagnetism. *Z. Phys*, 31(1):253–258, 1925.

M. L. Itz, J. Golle, S. Luttmann, S. R. Schweinberger, and J. M. Kaufmann. Dominance of texture over shape in facial identity processing is modulated by individual abilities. *British Journal of Psychology*, 108(2):369–396, 2017.

A. Jalali, P. Ravikumar, V. Vasuki, and S. Sanghavi. On learning discrete graphical models using group-sparse regularization. In *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 378–387, 2011.

E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106 (4):620–630, 1957.

G. K. Kamenev. Analysis of an algorithm for approximating convex bodies. *Computational Mathematics and Mathematical Physics*, 34(4):521–528, 1994.

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

J. D. Lee and T. J. Hastie. Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24(1):230–253, 2015.

J. D. Lee, Y. Sun, and J. E. Taylor. On model selection consistency of regularized M-estimators. *Electronic Journal of Statistics*, 9(1):608–642, 2015.

A. Löhne, B. Rudloff, and F. Ulus. Primal and dual approximation algorithms for convex vector optimization problems. *Journal of Global Optimization*, 60(4): 713–736, 2014.

C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan. Tensor robust principal component analysis: Exact recovery of corrupted low-rank tensors via convex optimization. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5249–5257, 2016.

C. Lu, J. Feng, Y. Chen, W. Liu, Z. Lin, and S. Yan. Tensor robust principal component analysis with a new tensor nuclear norm. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019.

S. Ma, L. Xue, and H. Zou. Alternating direction methods for latent variable Gaussian graphical model selection. *Neural computation*, 25(8):2172–2198, 2013.

M. Marsman, G. Maris, T. Bechger, and C. Glas. Bayesian inference for low-rank Ising networks. *Nature Scientific Reports*, 5(9050):1–7, 2015.

M. Marsman, D. Borsboom, J. Kruis, S. Epskamp, R. van Bork, L. J. Waldorp, H. L. J. van der Maas, and G. Maris. An introduction to network psychometrics: Relating Ising network models to Item Response Theory models. *Multivariate behavioral research*, 53(1):15–35, 2018.

M. McCoy and J. A. Tropp. Two proposals for robust PCA using semidefinite programming. *Electronic Journal of Statistics*, 5:1935–7524, 2011.

N. Meinshausen and P. Bühlmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

A. Mozeika, O. Dikmen, and J. Piili. Consistent inference of a general model using the pseudolikelihood method. *Physical Review E*, 90(1):010101, 2014.

P. Netrapalli, U. Niranjan, S. Sanghavi, A. Anandkumar, and P. Jain. Non-convex robust PCA. In *Advances in Neural Information Processing Systems*, pages 1107–1115, 2014.

F. Nussbaum and J. Giesen. Ising models with latent conditional Gaussian variables. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, pages 669–681, Chicago, Illinois, 22–24 Mar 2019a. PMLR.

F. Nussbaum and J. Giesen. Ising models with latent conditional Gaussian variables. Technical report, arXiv preprint arXiv:1901.09712, 2019b.

F. Nussbaum and J. Giesen. Disentangling direct and indirect interactions in polytomous item response theory models. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 2241–2247, 7 2020a. doi: 10.24963/ijcai.2020/310. URL `https://doi.org/10.24963/ijcai.2020/310`.

F. Nussbaum and J. Giesen. Pairwise sparse + low-rank models for variables of mixed type. *Journal of Multivariate Analysis*, 178:104601, 2020b. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2020.104601. URL `http://www.sciencedirect.com/science/article/pii/S0047259X19303756`.

F. Nussbaum and J. Giesen. Robust principal component analysis for generalized multi-view models. In *Conference on Uncertainty in Artificial Intelligence*. PMLR, 2021.

Open-Source Psychometrics Project. Open psychology data: Raw data from online personality tests. `https://openpsychometrics.org/_rawdata`, 2019.

R. Ostini and M. L. Nering. *Polytomous item response theory models*. Number 144 in Quantitative Applications in the Social Sciences. Sage, 2006.

K. Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *The Annals of Statistics*, 38(3): 1287–1319, 2010.

P. Ravikumar, M. J. Wainwright, G. Raskutti, and B. Yu. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electronic Journal of Statistics*, 5:935–980, 2011.

M. D. Sammel, L. M. Ryan, and J. M. Legler. Latent variable models for mixed discrete and continuous outcomes. *Journal of the Royal Statistical Society, Series B*, 59(3):667–678, 1997.

E. Schubert and A. Zimek. ELKI: A large open-source library for data analysis - ELKI release 0.7.5 "heidelberg". *CoRR*, abs/1902.03616, 2019. URL `http://arxiv.org/abs/1902.03616`.

J. Schur. Bemerkungen zur Theorie der beschränkten Bilinearformen mit unendlich vielen Veränderlichen. *Journal für die reine und Angewandte Mathematik*, 140: 1–28, 1911.

C. Spearman. "General intelligence," objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.

P. Sprechmann, A. M. Bronstein, and G. Sapiro. Learning efficient sparse and low rank models. *IEEE transactions on pattern analysis and machine intelligence*, 37 (9):1821–1833, 2015.

J. F. Sturm. Using sedumi 1.02, a matlab toolbox for optimization over symmetric cones. *Optimization methods and software*, 11(1-4):625–653, 1999.

S. Sun. A survey of multi-view machine learning. *Neural Computing and Applications*, 23(7-8):2031–2038, 2013.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

A. N. Tikhonov. On the stability of inverse problems. In *Dokl. Akad. Nauk SSSR*, volume 39, pages 195–198, 1943.

T. Tjaden, J. Bergner, J. Weniger, and V. Quaschning. Representative electrical load profiles of residential buildings in Germany with a temporal resolution of one second. *ResearchGate: Berlin, Germany*, 2015.

R. Tomioka and T. Suzuki. Convex tensor decomposition via structured schatten norm regularization. In *Advances in neural information processing systems*, pages 1331–1339, 2013.

A. Tsanas, M. A. Little, C. Fox, and L. O. Ramig. Objective automatic assessment of rehabilitative speech treatment in Parkinson's disease. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 22(1):181–190, 2014.

R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. Technical report, arXiv preprint arXiv:1011.3027, 2010.

M. J. Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Information Theory*, 55(5):2183–2202, 2009.

G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and its Applications*, 170:33–45, 1992.

M. Wedel and W. A. Kamakura. Factor analysis with (mixed) observed and latent variables in the exponential family. *Psychometrika*, 66(4):515–530, 2001.

J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma. Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization. In *Advances in neural information processing systems*, pages 2080–2088, 2009.

H. Xu, C. Caramanis, and S. Sanghavi. Robust PCA via outlier pursuit. In *Advances in Neural Information Processing Systems*, pages 2496–2504, 2010.

E. Yang, P. K. Ravikumar, G. I. Allen, and Z. Liu. On Poisson graphical models. In *Advances in neural information processing systems*, pages 1718–1726, 2013.

E. Yang, P. Ravikumar, G. I. Allen, and Z. Liu. Graphical models via univariate exponential family distributions. *The Journal of Machine Learning Research*, 16 (1):3813–3847, 2015.

X. Yu, T. Liu, X. Wang, and D. Tao. On compressing deep models by low rank and sparse decomposition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7370–7379, 2017.

R. Zhang, F. Nie, X. Li, and X. Wei. Feature selection with multi-view data: A survey. *Information Fusion*, 50:158–167, 2019.

Z. Zhang, G. Ely, S. Aeron, N. Hao, and M. Kilmer. Novel methods for multilinear data completion and de-noising based on tensor-SVD. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3842–3849, 2014.

J. Zhao, X. Xie, X. Xu, and S. Sun. Multi-view learning overview: Recent progress and new challenges. *Information Fusion*, 38:43–54, 2017.

F. Zhou, Q. Claire, and R. D. King. Predicting the Geographical Origin of Music. In *IEEE International Conference on Data Mining (ICDM)*, pages 1115–1120, 2014.

P. Zhou and J. Feng. Outlier-robust tensor PCA. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2263–2271, 2017.