cgmodsel: Technical Documentation

Frank Nussbaum, April 2022

1 Introduction

This is a technical report describing the algorithms used in the CGMODSEL Python package for estimating conditional Gaussian (CG) distributions from data. The repository is available under https://github.com/franknu/cgmodsel. This report complements the documentation of the interfaces provided by the module, which can be found here. Since the interfaces are documented elsewhere, here we focus only on the general characteristic of the package.

Organization of the package. There are two main components: (probabilistic) *models* and *solvers* to estimate the paramters from these models from data (using convex optimization).

Models are implemented as Python classes in the sub-package CGMODSEL.MODELS. Each model is characterized by a set of parameters. Methods for model instances can include

- MARGINALIZE: return a marginal model (keep only selected dimensions)
- CONDITIONALIZE: return a conditional model (must provide evidence)
- REPR_GRAPHICAL: display a graphical representation of the model
- GET_MEAN_PARAMS: convert canonical model parameters to *mean* parameters[1]
- SAVE: save a model in a file
- LOAD: load a model from a given file

Not all models instances in the package implement all of these methods yet.

Solvers are also implemented as Python classes, see Section 3 for an introduction to the algorithms (most implemented solvers are variants of the *Alternating Direction Method of Multipliers (ADMM)* [1]). Solvers are directly situated in the CGMODSEL package. Typical methods are:

- DROP_DATA: deposit data in the instance
- SET_REGULARIZATION_PARAMS: specify hyper-/regularization parameters for corresponding (convex) optimization problem.
- SOLVE: estimate parameters by solving corresponding (convex) optimization problem
- GET_PARAMS: return parameters as instance of the corresponding model class

2 Pairwise CG models

A comprehensive introduction to pairwise CG models, particularly ones with sparse + low-rank decompositions can be found in the dissertation [5].

[1] We use canonical parameterization as the default parameterization because it relates more directly to graphical models, see also Section 2.

Setup and notation. We consider the general CG (conditional Gaussian) framework, that is, observed dimensions can be of categorical/discrete or quantitative type. Let d denote the number of categorical variables and q denote the number of continuous, quantitative (conditional Gaussian) variables. By L_r we denote the number of levels of the r-th discrete variable. By $L_{\mathrm{tot}} = \sum_{r=1}^{d} L_r$ we denote the total number of discrete levels.

Generally, we assume the sample space to be $\mathcal{X} \times \mathcal{Y}$ with the discrete-label space $\mathcal{X} = [L_1] \times [L_2] \times \cdots \times [L_d]$ using the definition $[a] = \{1, \ldots, a\}$ for a natural number a. Moreover, $\mathcal{Y} = \mathbb{R}^q$ is the sample space of the continuous, conditional Gaussian variables.

For $a \in \mathbb{N}$, we denote the set of symmetric $(a \times a)$ matrices by $\mathrm{Sym}(a)$.

2.1 Probability model and parameters

A general pairwise conditional Gaussian (CG) model in canonical parameterization is given by (up to normalization), for $(x, y) \in \mathcal{X} \times \mathcal{Y}$

$$\begin{aligned} p(x,y) &\propto \exp\left(u^\top \overline{x} + \frac{1}{2}\overline{x}^\top Q \overline{x} + (R\overline{x})^\top y + \alpha^\top y - \frac{1}{2} y^\top \Lambda y\right) \\ &= \exp\left(\frac{1}{2}\begin{pmatrix}\overline{x} \\ y\end{pmatrix}^\top \begin{pmatrix} Q & R^\top \\ R & -\Lambda\end{pmatrix}\begin{pmatrix}\overline{x} \\ y\end{pmatrix} + u^\top \overline{x} + \alpha^\top y\right) \\ &= \exp\left(\frac{1}{2}\begin{pmatrix}\overline{x} \\ y\end{pmatrix}^\top \begin{pmatrix} Q + 2\,\mathrm{diag}(u) & R^\top \\ R & -\Lambda\end{pmatrix}\begin{pmatrix}\overline{x} \\ y\end{pmatrix} + \alpha^\top y\right). \end{aligned}$$

Here, for a vector $x \in \mathcal{X}$ we denoted the concatenated *indicator* vectors of the values of the discrete variables by

$$\begin{aligned} \overline{x} &= (\overline{x}_1, \ldots, \overline{x}_d) \in \mathbb{R}^{L_{\mathrm{tot}}} \\ \overline{x}_r &= (\{\mathbb{1}[x_r = 1]\}, \ldots, \{\mathbb{1}[x_r = L_r]\}) \in \mathbb{R}^{L_r}, \qquad r = 1, \ldots, d. \end{aligned}$$

Moreover, the interaction parameters are as follows

- $Q \in \mathrm{Sym}(L_{\mathrm{tot}})$ contains the interaction parameters $q_{r:k,j:l} = q_{rj}(k,l)$ between discrete variables x_r and x_j, respectively with values k and l (Q has zero 'block diagonal', that is, the elements $q_{rr}(\cdot,\cdot)$ are zero, respectively for all $r \in [d]$),
- $u \in \mathbb{R}^{L_{\mathrm{tot}}}$ are the univariate discrete parameters with entries $q_{r:k} = q_r(k)$,
- $R \in \mathbb{R}^{q \times L_{\mathrm{tot}}}$ contains the interaction parameters $\rho_{s,r:k} = \rho_{sr}(k)$ between conditional Gaussians y_s and discrete variables x_r with value k,
- $\Lambda \in \mathrm{Sym}(q)$ is the precision matrix of the conditional Gaussian variables (must be positive definite),
- $\alpha \in \mathbb{R}^q$ are the univariate continuous parameters.

For the pairwise interaction parameter matrix we write

$$\Theta = \begin{pmatrix} Q + 2\,\mathrm{diag}(u) & R^\top \\ R & -\Lambda\end{pmatrix} \in \mathrm{Sym}(L_{\mathrm{tot}} + q).$$

Here, we added the discrete univariate parameters to the diagonal of the discrete interaction parameter matrix Q. Using entries of the components of the parameter matrix, the pairwise CG density can also be written as

$$\begin{aligned}
p(x,y) &\propto \exp\left(\sum_{r,j=1}^{d} q_{rj}(x_r,x_j) + \sum_{r=1}^{d}\sum_{s=1}^{q} \rho_{sr}(x_r) - \frac{1}{2}\sum_{s,t=1}^{q} \lambda_{st} y_s y_t + \sum_{r=1}^{d} u_r(x_r) + \sum_{s=1}^{q} \alpha_s y_s\right) \\
&= \exp\Big(\sum_{r,j=1}^{d}\sum_{k=1}^{L_r}\sum_{l=1}^{L_j} q_{r:k,j:l}\mathbb{1}[x_r = k, x_j = l] + \sum_{r=1}^{d}\sum_{s=1}^{q}\sum_{k=1}^{L_r} \rho_{s,r:k}\mathbb{1}[x_r = k] \\
&\qquad \ldots - \frac{1}{2}\sum_{s,t=1}^{q} \lambda_{st} y_s y_t + \sum_{r=1}^{d}\sum_{k=1}^{L_r} u_{r:k}\mathbb{1}[x_r = k] + \sum_{s=1}^{q} \alpha_s y_s\Big).
\end{aligned}$$

Likelihoods. Let n data points $(x^{(i)}, y^{(i)})$ for $i = 1, \ldots, n$ be given. The *negative* standard log-likelihood is defined as

$$\ell(\Theta, \alpha) = -\sum_{i=1}^{n} \log p(x^{(i)}, y^{(i)}).$$

This standard version requires computation of the normalization constant. Especially in the presence of discrete variables this amounts to computing large sums which easily becomes computationally intractable. A more tractable alternative is the *negative* pseudo log-likelihood given by

$$\ell_p(\Theta, \alpha) = -\sum_{i=1}^{n}\left(\sum_{r} \log p(x_r = x_r^{(i)} | x_{-r}^{(i)}, y^{(i)}) + \sum_{s} \log p(y_s = y_s^{(i)} | x^{(i)}, y_{-s}^{(i)})\right). \tag{1}$$

More details on its construction can be found in Appendix A and in the dissertation [5].

Note that we use negative versions of the respective likelihoods for convenience (to be able 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.

Identifiability (unique parametrizations). The class of pairwise CG distributions is non-identifiable, that is, there exist different parameter configurations that yield the same distribution.

The class can be made identifiable by requiring that for all $r \in [d]$ and $k \in [L_r]$

$$q_r(0) = 0, q_{rj}(k, 0) = 0, q_{rj}(0, k) = 0, \text{ and } \quad \rho_{sr}(0) = 0. \tag{2}$$

Effectively this means that for each discrete variable the 0-th is excluded from interacting.

Note that there are other ways to obtain identifiable parameter classes. [3] use sparse regularization, see the next section, since then implicitly the model with the lowest sparse norm is preferred.

2.2 Special cases

Gaussians. The purely Gaussian model has density (using canonical parameters)

$$p(y) = p(y; \Lambda, \alpha) = (2\pi)^{-q/2}\det(\Lambda)^{1/2}\exp\left(\alpha^\top y - \frac{1}{2}y^\top \Lambda y\right).$$

Assume that $y^{(1)}, \ldots, y^{(n)} \in \mathbb{R}^q$ are n zero-mean data points, that is, they have density

$$\begin{aligned}
p(y) = p(y; \Lambda) &= (2\pi)^{-q/2}\det(\Lambda)^{1/2}\exp\left(-\frac{1}{2}y^\top \Lambda y\right) \\
&= (2\pi)^{-q/2}\det(\Lambda)^{1/2}\exp\left(-\frac{1}{2}\left\langle \Lambda, yy^\top\right\rangle\right).
\end{aligned}$$

Then, for this zero-mean model the standard Gaussian log-likelihood is given by

$$\begin{aligned}\ell(\Lambda) &= \log\left(\prod_{i=1}^{n} p(y^{(i)};\Lambda)\right) = \sum_{i=1}^{n} \log p(y^{(i)};\Lambda)\\ &= \sum_{i=1}^{n}\left(-\frac{q}{2}\log(2\pi) + \frac{1}{2}\log\det(\Lambda) - \frac{1}{2}\left\langle \Lambda, y^{(i)}\left[y^{(i)}\right]^{\top}\right\rangle\right)\\ &= -\frac{nq}{2}\log(2\pi) + \frac{n}{2}\log\det(\Lambda) - \frac{n}{2}\left\langle \Lambda, \Sigma_0\right\rangle,\end{aligned}$$

where $\Sigma_0 = \frac{1}{n}\sum_{i=1}^{n} y^{(i)}\left[y^{(i)}\right]^{\top}$ is the empirical covariance matrix.

With some abuse of notation, for optimization the *negative* scaled log-likelihood, ignoring constant parts that do not depend on Λ, is used:

$$\ell(\Lambda) = -\log\det(\Lambda) + \left\langle \Lambda, \Sigma_0\right\rangle. \tag{3}$$

Multivariate binary. The multivariate binary pairwise model for vectors $x \in \{0,1\}^d$ has density

$$p(x) \propto \exp\left(u^{\top}x + \frac{1}{2}x^{\top}Qx\right) = \exp\left(\frac{1}{2}x^{\top}\left(Q + 2\operatorname{diag}(u)\right)x\right),$$

where now $u \in \mathbb{R}^d$ and $Q \in \operatorname{Sym}(d)$, omitting the parameters set to zero by the identifiability condition (2).

The node conditionals are logistic models

$$p(x_r|x_{-r}, y) = \frac{\exp\left(u_r x_r + \sum_{r\neq j} q_{rj}x_r x_j\right)}{1 + \exp\left(u_r + \sum_{r\neq j} q_{rj}x_j\right)}, \qquad r = 1,\dots,d.$$

This is a special case of the general form (16) of discrete node conditionals. Now, based on n observations $x^{(1)},\dots,x^{(n)} \in \{0,1\}^d$, the *negative* pseudo log-likelihood is given by

$$\ell_p(Q,u) = -\sum_{i=1}^{n}\left(\sum_{r}\log p(x_r = x_r^{(i)}|x_{-r}^{(i)}, y^{(i)})\right). \tag{4}$$

2.3 Pairwise sparse model

In this section, we omit univariate parameters for clarity (they can be trivially added to the model). For sparse models, the pairwise interaction parameter matrix

$$S = \begin{pmatrix} Q & R^{\top} \\ R & -\Lambda \end{pmatrix}$$

is assumed to be (group-)sparse in the sense that the associated graphical model has few edges. The following parameter groups are associated with the edges between

- two continuous variables y_s and y_t, $s \neq t$: parameters β_{st},
- continuous variable y_s and discrete variable x_r: parameters $\rho_{sr} = \{\rho_{sr}(k)\}_{k\in[L_r]}$,
- two discrete variables x_r and x_j, $r \neq j$: parameters $Q_{rj} = \{q_{rj}(k,l)\}_{k\in[L_r],l\in[L_j]}$.

This leads to the group-sparsity term (off-diagonal $\ell_{2,1}$-norm)

$$\|S\|_{2,1} = \sum_{s\neq t}|\beta_{st}| + 2\sum_{r\in[d],s\in[q]}\|\rho_{sr}\|_2 + \sum_{r\neq q}\|Q_{rj}\|_F.$$

In the special case that all discrete variables are binary, the identifiability constraints (2) imply that each parameter group contains only one parameter. In this case, the $\ell_{2,1}$-norm reduces to the ℓ_1-norm.

In the general case, a sparse pairwise CG model (omitting univariate parameters) can be learned via the following convex optimization problem

$$\min_{S} \quad \ell(S) + \lambda \|S\|_{2,1} \quad \text{s.t.} \quad \Lambda[S] \succ 0 \ . \tag{S}$$

Here, ℓ is some likelihood (typically either standard negative log-likelihood or negative pseudo log-likelihood, see Appendix A for the construction) and $\lambda > 0$ is a regularization parameter. Furthermore, $\Lambda[S]$ selects the continuous-continuous interaction parameters (precision matrix) from the bottom right corner of the pairwise interaction parameter matrix S. The constraint $\Lambda[S] \succ 0$ thereby ensures normalizability of the pairwise CG density.

2.4 Pairwise sparse + low-rank models

This is an extension of sparse graphical models from the previous section, originally proposed for Gaussians in [2]. Now, the interaction parameter matrix decomposes into a sparse and a low-rank component. The low-rank component represents spurious indirect interaction in between the observed variables that are caused by a small number of latent conditional Gaussian variables, see also [6]. See the dissertation [5] for details in the context of CG distributions.

A (group-)sparse + low-rank model is learned using the following convex optimization problem

$$\min_{S,\, L} \quad \ell(S + L) + \lambda \|S\|_{2,1} + \rho\,\mathrm{tr}(L) \quad \text{s.t.} \quad L \succeq 0, \ \text{and } \Lambda[S + L] \succ 0 \ , \tag{SL}$$

where the trace or nuclear norm $\mathrm{tr}(L)$ induces low rank on L. The regularization parameters $\lambda, \rho > 0$ are used as trade-off parameters.

3 ADMM-like algorithms

In this section we formally describe the ADMM-algorithms implemented in the package.

A general note in advance: To keep the descriptions of the algorithms concise we omit lower-order sufficient statistics (so no univariate parameters). In the implementation they can be incorporated in the respective likelihood optimization steps.

3.1 Introduction to ADMM

This is based on the survey [1]. Here, we only consider a special case suitable for our problems. Consider the optimization problem

$$\begin{aligned} \min_{X,Z} \quad & f(X) + g(Z) \\ \text{s.t.} \quad & X = Z, \end{aligned} \tag{5}$$

The augmented Lagrangian for this problem reads as

$$\mathcal{L}(X, Z, \Phi) = f(X) + g(Z) - \langle \Phi, X - Z \rangle + \frac{1}{2\mu} \|X - Z\|_F^2 \,. \tag{6}$$

Note the difference compared to [1] who write $\rho = \mu^{-1}$ and conceptually there dual variables $y = -\Phi$. We use capital letters for the variables here, since our ADMM algorithms are for objectives dependent on matrices. Now, the ADMM updates are

$$\begin{cases} X^{k+1} & = \arg\min_X \ f(X) - \left\langle \Phi^k, X - Z^k \right\rangle + \frac{1}{2\mu} \left\| X - Z^k \right\|_F^2 , \\ Z^{k+1} & = \arg\min_Z \ g(Z) - \left\langle \Phi^k, X^{k+1} - Z \right\rangle + \frac{1}{2\mu} \left\| X^{k+1} - Z \right\|_F^2 , \\ \Phi^{k+1} & = \Phi^k - \mu^{-1}(X^{k+1} - Z^{k+1}) \end{cases}$$

which is equivalent to

$$\begin{cases} X^{k+1} &= \arg\min_X \ f(X) + \frac{1}{2\mu} \left\| X - Z^k - \mu\Phi^k \right\|_F^2, \\ Z^{k+1} &= \arg\min_Z \ g(Z) + \frac{1}{2\mu} \left\| X^{k+1} - Z - \mu\Phi^k \right\|_F^2 \\ \Phi^{k+1} &= \Phi^k - \mu^{-1}(X^{k+1} - Z^{k+1}) \end{cases} \tag{7}$$

The first two updates are proximal operators of the functions f and g. In the next section, we discuss the most important proximal operators that are relevant for (latent variable) graphical models.

3.2 Proximal operators

The proximal operator of a scalar convex function f on some vector space is defined by

$$\mathrm{prox}(f, \kappa, v) = \arg\min_x f(x) + \frac{1}{2\kappa}\|x - v\|_2^2,$$

where $\kappa > 0$ defines the strength how much deviation of the solution from a vector v is penalized.

Here we provide a list of proximal operators that are required by ADMM-algorithms throughout this section.

Gaussian likelihood. The zero-mean Gaussian negative log-likelihood is given by

$$\ell(\Theta) = -\log\det\Theta + \langle\Theta, \Sigma_0\rangle + \mathcal{I}(\Theta \succ 0)$$

with empirical covariance matrix Σ_0, see (3). The corresponding proximal operator

$$\mathrm{prox}(\ell, \kappa, Z) = \arg\min_\Theta \ell(\Theta) + \frac{1}{2\kappa}\|\Theta - Z\|_F^2$$

has an analytical solution which can be derived from the first-order optimality condition

$$-\Theta^{-1} + \Sigma_0 + \frac{1}{\kappa}(\Theta - Z) \overset{!}{=} 0$$

or equivalently

$$\Theta \overset{!}{=} \kappa\Theta^{-1} - (\kappa\Sigma_0 - Z).$$

Let $U\,\mathrm{diag}(\sigma)U^\top$ be a singular value decomposition of $\kappa\Sigma_0 - Z$. Assume that the solution of the stationary equation has the form $\Theta = U\,\mathrm{diag}(\gamma)U^\top$. Then, since

$$\Theta = U\,\mathrm{diag}(\gamma)U^\top \overset{!}{=} \kappa\Theta^{-1} - (\kappa\Sigma_0 - Z) = \kappa U\,\mathrm{diag}(\gamma^{-1})U^\top - U\,\mathrm{diag}(\sigma)U^\top$$

it must hold $\gamma = \kappa\gamma^{-1} - \sigma$. This is essentially a quadratic equation in γ. We solve it componentwise ($\gamma_i^2 + \sigma_i\gamma_i - \kappa = 0$), which yields

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

Note that only this positive solution gives a positive definite parameter matrix Θ. We have shown that

$$\mathrm{prox}(\ell, \kappa, Z) = U\,\mathrm{diag}(\gamma)U^\top, \tag{Prox-GLH}$$

see also [4].

Pseudo likelihood. The proximal operator for the negative pseudo log-likelihood ℓ_p with zero univariate parameters as defined in (1) does in general not have a closed form analytical solution. The problem

$$\mathrm{prox}(\ell_p, \kappa, Z) = \arg\min_\Theta \ell_p(\Theta) + \frac{1}{2\kappa}\|\Theta - Z\|_F^2$$

needs to be solved using iterative solver. Since the objective is differentiable, gradient descent algorithms such as L-BFGS-B can be used (alternatively use GENO).

Sparse norms. For the matrix ℓ_1-norm it holds

$$\operatorname{prox}(\|\cdot\|_1, \kappa, Z) = \arg\min_S \|S\|_1 + \frac{1}{2\kappa}\|S - Z\|_F^2 = \operatorname{Shrink}(Z, \kappa), \tag{Prox-ℓ_1}$$

where for all entries (i, j) the element-wise soft-threshold is defined by

$$[\operatorname{Shrink}(Z, \kappa)]_{ij} = \operatorname{sign}(Z_{ij}) \cdot \max\{|Z_{ij}| - \kappa, 0\}$$

Similarly, for the $\ell_{2,1}$-norm with groups $g \in \mathcal{G}$

$$\operatorname{prox}(\|\cdot\|_{2,1}, \kappa, Z) = \arg\min_S \|S\|_{2,1} + \frac{1}{2\kappa}\|S - Z\|_F^2 = \operatorname{gShrink}(Z, \kappa), \tag{Prox-$\ell_{2,1}$}$$

where

$$[\operatorname{gShrink}(Z, \kappa)]_g = \begin{cases} Z_g(1 - \frac{\kappa}{\|Z_g\|_2}), & \|Z_g\|_2 > \kappa \\ 0, & \text{else} \end{cases}.$$

Low-rank norm(s). For the matrix nuclear norm (or Schatten S_1-norm) it holds

$$\operatorname{prox}(\|\cdot\|_*, \kappa, Z) = \arg\min_L \|L\|_* + \frac{1}{2\kappa}\|L - Z\|_F^2 = \operatorname{sShrink}(Z, \kappa), \tag{Prox-S_1}$$

where

$$\operatorname{sShrink}(Z, \kappa) = U \operatorname{Shrink}(\operatorname{diag}(\sigma), \kappa) U^\top$$

provided a singular value decomposition $Z = U \operatorname{diag}(\sigma) U^\top$ of Z.

3.3 Gaussian unregularized problem

→ can use the class CGMODSEL.ADMMGAUSSIANPW, solve with regularization parameter zero (see below).

The unregularized Gaussian likelihood problem is

$$\min_\Theta \quad \ell(\Theta) \tag{8}$$

where

$$\ell(\Theta) = -\log\det\Theta + \langle\Theta, \Sigma_0\rangle + \mathcal{I}(\Theta \succ 0)$$

with empirical covariance matrix Σ_0. If the empirical covariance matrix is non-degenerate, the problem has the analytical solution $\Theta = \Sigma_0^{-1}$ based on the first order optimality condition. Otherwise an update idea is

$$\begin{cases} \Theta^{k+1} & = \arg\min_\Theta \ell(\Theta) + \frac{1}{2\mu}\left\|\Theta - \Theta^k\right\|_F^2 = \operatorname{prox}(\ell, \mu, \Theta^k) = U \operatorname{diag}(\gamma) U^\top, \end{cases}$$

where $U \operatorname{diag}(\sigma) U^\top$ is a singular value decomposition of $\mu\Sigma_0 - \Theta^k$, see (Prox-GLH).

3.4 Graphical Lasso with ADMM

3.4.1 Gaussian Graphical Lasso using standard likelihood

→ Implemented in the class CGMODSEL.ADMMGAUSSIANPW.

An ADMM formulation is

$$\min_{\Theta, S} \quad \ell(\Theta) + \varphi(S) \quad \text{s.t.} \quad \Theta = S \tag{9}$$

where

$$\ell(\Theta) = -\log\det\Theta + \langle\Theta, \Sigma_0\rangle + \mathcal{I}(\Theta \succ 0)$$

is the zero-mean Gaussian negative log-likelihood as in (3) with empirical covariance matrix Σ_0, next $\mathcal{I}$ is the 0-∞ indicator function of the positive definite cone, and

$$\varphi(S) = \alpha\|S\|_1.$$

Let Φ be the dual variables for the equality constraint. The augmented Lagrangian for the ADMM problem then reads as

$$\mathcal{L}(\Theta, S, \Phi) = \ell(\Theta) + \varphi(S) - \langle\Phi, \Theta - S\rangle + \frac{1}{2\mu}\left\|\Theta - S\right\|_F^2, \tag{10}$$

where $\mu > 0$ is the ADMM parameter that controls the strength of the additional quadratic penalty for violation of the equality constraint. The dual function is

$$g(\Phi) = \min_{\Theta,\, S}\ \ell(\Theta) + \varphi(S) - \langle\Phi, \Theta - S\rangle + \frac{1}{2\mu}\left\|\Theta - S\right\|_F^2. \tag{11}$$

ADMM proceeds by conducting the minimization with respect to Θ and S in sequential order, followed by a gradient step for the dual problem (that is, maximizing the dual function g), see [1]. Hence, the ADMM updates are

$$\begin{cases} \Theta^{k+1} & = \arg\min_\Theta\ \ell(\Theta) - \left\langle\Phi^k, \Theta - S^k\right\rangle + \frac{1}{2\mu}\left\|\Theta - S^k\right\|_F^2, \\ S^{k+1} & = \arg\min_S\ \varphi(S) - \left\langle\Phi^k, \Theta^{k+1} - S\right\rangle + \frac{1}{2\mu}\left\|\Theta^{k+1} - S\right\|_F^2, \\ \Phi^{k+1} & = \Phi^k - \mu^{-1}(\Theta^{k+1} - S^{k+1}) \end{cases}$$

which is equivalent to

$$\begin{cases} \Theta^{k+1} & = \arg\min_\Theta\ \ell(\Theta) + \frac{1}{2\mu}\left\|\Theta - S^k - \mu\Phi^k\right\|_F^2, \\ S^{k+1} & = \arg\min_S\ \varphi(S) + \frac{1}{2\mu}\left\|\Theta^{k+1} - S - \mu\Phi^k\right\|_F^2 \\ \Phi^{k+1} & = \Phi^k - \mu^{-1}(\Theta^{k+1} - S^{k+1}) \end{cases} \tag{12}$$

The first step has the analytical solution (Prox-GLH) with $Z = S + \mu\Phi^k$ and $\kappa = \mu$. The second step is the soft-thresholding operation (Prox-ℓ_1) with $Z = \Theta^{k+1} - \mu\Phi^k$ and $\kappa = \alpha\mu$.

Concerning univariate parameters. Observe that minimizing the negative log-likelihood

$$\ell(\Theta, \alpha) = -\log\det\Theta + \langle\Theta, \Sigma_0\rangle - \alpha^\top\mu_0 + \mathcal{I}(\Theta \succ 0)$$

including univariate parameters is an unbounded problem whenever the empirical mean $\mu_0 = \frac{1}{n}\sum_{i=1}^n y^{(i)}$ is non-zero. Therefore, the data should be centered beforehand by subtracting μ_0 from each data point. Afterwards a model with $\alpha = 0$ should be learned.

3.4.2 CG Graphical Lasso using pseudo likelihood

→ Implemented in the class CGMODSEL.ADMMCGAUSSIANPW.

For CG models, it is more convenient to use pseudo likelihood. The ADMM problem (with zero univariate parameters) is

$$\min_{\Theta,\, S}\quad \ell_p(\Theta) + \varphi(S) \quad \text{s.t.} \quad \Theta = S \tag{13}$$

where, compare (1),

$$\ell_p(\Theta) = -\sum_{i=1}^n\left(\sum_r \log p(x_r^{(i)}|x_{-r}^{(i)}, y^{(i)}; \Theta) + \sum_s \log p(y_s^{(i)}|x^{(i)}, y_{-s}^{(i)}; \Theta)\right) + \mathcal{I}(\Lambda[\Theta] \succ 0)$$

is the negative pseudo-loglikelihood for n data points.

The ADMM algorithm is the same as above, but the first update now becomes the proximal mapping of the negative pseudo log-likelihood given by

$$\Theta^{k+1} = \arg\min_{\Theta} \ell_p(\Theta) + \frac{1}{2\mu} \left\|\Theta - S^k - \mu\Phi^k\right\|_F^2 .$$

This needs to be solved with an iterative solver. Unfortunately, this limits the performance of ADMM algorithms for pseudo likelihood problems.

Concerning univariate parameters. In a model where the univariate parameters u and α are allowed to be non-zero, the first update just becomes

$$\Theta^{k+1} = \arg\min_{\Theta,\, \alpha} \ell_p(\Theta, \alpha) + \frac{1}{2\mu}\|\Theta - S^k - \mu\Phi^k\|_F^2,$$

where the univariate discrete parameters u are on the part of the diagonal of Θ that is associated with discrete variables. Note that in contrast to the Gaussian likelihood, minimization w.r.t. α is not an unbounded problem for the pseudo log-likelihood, see also Appendix A.

3.5 CG S+L with proximal-gradient based ADMM

→ Implemented in the class cgmodsel.AdmmCGaussianSL (uses pseudo-likelihood). A version using likelihood for purely Gaussian models is implemented in the class cgmodsel.AdmmGaussianSL.

Let us first consider a reformulation of Problem SL

$$\arg\min_{\Theta \succ 0,\, W} \quad \ell(\Theta) + \varphi(W) \quad \text{s.t.} \quad \Theta = [I, I]W \tag{14}$$

where we grouped $W = (S, L)$ into one variable and let

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

Let $\mu > 0$ and let Φ be the dual variables for the constraint. The ADMM updates are

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

which is equivalent to

$$\begin{cases} \Theta^{k+1} & = \arg\min_{\Theta \succ 0} \ \ell(\Theta) + \frac{1}{2\mu} \left\|\Theta - [I, I]W^k - \mu\Phi^k\right\|_F^2 , \\ W^{k+1} & = \arg\min_{W} \ \varphi(W) + \frac{1}{2\mu} \left\|\Theta^{k+1} - [I, I]W - \mu\Phi^k\right\|_F^2 \\ \Phi^{k+1} & = \Phi^k - \mu^{-1}(\Theta^{k+1} - [I, I]W^{k+1}) \end{cases} \tag{15}$$

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

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

where $Z = [I, I]W^k + \mu\Phi^k$. For purely Gaussian models with standard likelihood, the solution is once again given by (Prox-GLH) with $\kappa = \mu$. As mentioned earlier, in the presence of binary variables neither likelihood nor pseudo likelihood proximal mappings have analytical solutions. Hence, an iterative optimization algorithm needs to be applied.

The second update. In the second problem of (12), the components of W are coupled in the quadratic Frobenius-norm term. With this coupling the proximal operator for W is hard to solve. Instead it has been suggested in [4] to solve a step of a proximal-gradient method, that is, for $\tau > 0$ one solves

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

The good news is that now the components S and L are separable. Consequently the proximal-gradient step reduces to solving two proximal operators, namely the one of the ℓ_1-norm

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

compare (Prox-$\ell_{2,1}$), and the one of the nuclear norm

$$\begin{aligned} L^{k+1} &= \arg\min_L \beta \operatorname{tr}(L) + \mathcal{I}(L \succeq 0) + \frac{1}{2\mu\tau} \left\| L - \left(L^k + \tau G^k\right) \right\|_F^2 L^{k+1} \\ &= \text{sShrink}(L^k + \tau G^k, \beta\mu\tau), \end{aligned}$$

compare (Prox-S_1). Here, we used $G^k = \Theta^{k+1} - S^k - L^k - \mu\Phi^k$ (TODO: is the negative partial gradient, some details on the proxgrad method).

References

[1] Stephen Boyd, Neal Parikh, Eric Chu, Borja Peleato, Jonathan Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

[2] Venkat Chandrasekaran, Pablo A. Parrilo, and Alan S. Willsky. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012.

[3] Jason D. Lee and Trevor J. Hastie. Learning the structure of mixed graphical models. *Journal of Computational and Graphical Statistics*, 24(1):230–253, 2015.

[4] Shiqian Ma, Lingzhou Xue, and Hui Zou. Alternating direction methods for latent variable Gaussian graphical model selection. *Neural computation*, 25(8):2172–2198, 2013.

[5] Frank Nussbaum. *Models with low-rank and group-sparse components and their recovery via convex optimization.* PhD thesis, 2021.

[6] Frank Nussbaum and Joachim Giesen. Ising models with latent conditional gaussian variables. In *Algorithmic Learning Theory*, pages 669–681. PMLR, 2019.

A Technical Pseudo-likelihood construction for pairwise CG models

Notation. For any natural number a we let $[a] = \{1, \ldots, a\}$. In empirical settings, n is the number of samples.

Indexing. We use r, j as typical indices for discrete variables and s, t as indices for continuous variables. For a dimension of size L_{tot} we use the index notation $r : k$ where $r \in [d]$ and $k \in [L_r]$. For matrices we use $\cdot$ to denote a slice, for example, for a matrix $Q \in \mathbb{R}^{L_{\text{tot}} \times L_{\text{tot}}}$ the $r : k$-th column is denoted by $Q_{\cdot, r:k}$.

A.1 Parameters and data

In Section 2, the parameter groups Q, u, R, Λ, α of a pairwise CG model have been specified. A useful trick in the implementation to implement the positive definiteness constraints on the precision matrix Λ is to write it as $\Lambda = FF^\top \succeq 0$, where

- $F \in \mathbb{R}^{q \times q}$ is a root of Λ.

Strictly speaking, this does only enforce positive *semi*-definiteness. Moreover, it is useful in practice to split the precision matrix into $\Lambda = B + \operatorname{diag}(\beta)$ with

- $B \in \operatorname{Sym}(q)$ holds the quantitative-quantitative interaction parameters β_{st} (with zero diagonal).
- $\beta \in \mathbb{R}^q_{>0}$ are the univariate squared continuous parameters $\beta_s = \beta_{ss}$ (that is, the diagonal of the conditional Gaussian precision matrix)

This allows easier construction of the (pseudo-likelihood) objective terms. This is particularly useful since the gradients w.r.t. diagonal and off-diagonal parameters of the precision matrix are calculated differently.

The implementation optimizes only the upper right triangle of parameters of Q. These are copied down to the lower half before each objective evaluation.

Data. We denote the continuous data by $Y \in \mathbb{R}^{n \times q}$ and we denote the discrete indicator data by $D \in \mathbb{R}^{n \times L_{\text{tot}}}$. Here, rows respectively correspond to samples. This is motivated by the typical representation of data in form of tables.

A.2 Discrete node conditional

For the node conditional of the r-th discrete variable only the parameters that are associated with x_r taking on the value k are relevant. Indeed, the discrete node conditional of a pairwise model has the form of a logistic probability

$$p(x_r = k | x_{-r}, y) = \frac{\exp\Big(q_r(k) + \sum_{r \neq j} q_{rj}(k, x_j) + \sum_s \rho_{sr}(k)\, y_s\Big)}{\sum_{l \in [L_r]} \exp\Big(q_r(l) + \sum_{r \neq j} q_{rj}(l, x_j) + \sum_s \rho_{sr}(l)\, y_s\Big)}. \tag{16}$$

Let $W = \mathbb{1}_n u^\top + DQ + YR \in \mathbb{R}^{n \times L_{tot}}$ and W_r denotes the slice of W that contains the columns from $r{:}1$ to $r{:}L_r$ and likewise by D_r the corresponding slice of the discrete data matrix D, that is,

$$W = \big(W_1 \mid W_2 \mid \cdots \mid W_d\big) \quad \text{and} \quad D = \big(D_1 \mid D_2 \mid \cdots \mid D_d\big).$$

Observe that the elements of W_r exactly reproduce the sums under the exponential function in (16), that is,

$$\begin{aligned} W_{i,r:k} &= u_{r:k} + D_{i,\cdot} Q_{\cdot,r:k} + y^{(i)} R_{\cdot,r} \\ &= q_r(k) + \sum_{r \neq j} q_{rj}(k, x_j^{(i)}) + \sum_s \rho_{s,r}(k)\, y_s^{(i)}, \end{aligned}$$

recalling that Q has zero block diagonal. Now, the log-likelihood for the r-th discrete variable is given by

$$\begin{aligned}
l_r &= -\sum_{i=1}^{n} \log p(x_r = x_r^{(i)} | x_{-r}^{(i)}, y^{(i)}) \\
&= -\left[\log\left(\left(\frac{\exp(W_r)}{\exp(W_r)\mathbb{1}_{L_r}} \odot D_r\right)\mathbb{1}_{L_r}\right)\right]^\top \mathbb{1}_n \\
&= -\left[\log\left(\frac{(\exp(W_r)\odot D_r)\,\mathbb{1}_{L_r}}{\exp(W_r)\mathbb{1}_{L_r}}\right)\right]^\top \mathbb{1}_n \\
&= -\left[(W_r \odot D_r)\mathbb{1}_{L_r} - \log\left(\exp(W_r)\mathbb{1}_{L_r}\right)\right]^\top \mathbb{1}_n
\end{aligned}$$

where in combination with the summation over the rows using $\mathbb{1}_{L_r}$, the element-wise multiplication with D_r, denoted by $\odot$, selects the correct term for the enumerator of (16) for each data point (recall that in each row of D_r all except one entry are zero). The last line can be implemented using a numerical stable version of *logsumexp*.

Gradients. Write $v^{(i)} = (1; d^{(i)}; y^{(i)})$ for $i \in [n]$ and let

$$\theta_{rk} = (u_{r:k}; Q_{\cdot,r:k}; R_{\cdot,r:k}) \in \mathbb{R}^{1+L_{\mathrm{tot}}+q}.$$

Then, the log node conditional, compare (16), can be written as

$$\log p(x_r = k | v_r^{(i)}) = \theta_{rk}^\top v^{(i)} - \log\left\{\sum_{l=1}^{L_r} \exp(\theta_{rl}^\top v^{(i)})\right\}.$$

We consider the gradient of the negative loglikelihood

$$l_r = -\sum_{i=1}^{n} \log p(x_r = x_r^{(i)} | x_{-r}^{(i)}, y^{(i)})$$

of the r-th node conditional w.r.t. the parameter vector θ_{rj}. It is given by

$$\begin{aligned}
\nabla_{\theta_{rj}} l_r &= \nabla_{\theta_{rj}}\left(-\sum_{i=1}^{n}\sum_{k=1}^{L_r} \mathbb{1}[x_r^{(i)} = k] \log p(x_r = k | v^{(i)})\right) \\
&= \nabla_{\theta_{rj}}\left(\sum_{i=1}^{n}\sum_{k=1}^{L_r} \mathbb{1}[x_r^{(i)} = k]\left(\log\left\{\sum_{l=1}^{L_r}\exp(\theta_{rl}^\top v^{(i)})\right\} - \theta_{rk}^\top v^{(i)}\right)\right) \\
&= \sum_{i=1}^{n}\Big(p(x_r = j | v^{(i)}) - \mathbb{1}[x_r^{(i)} = j]\Big)\, v^{(i)} \\
&= \sum_{i=1}^{n}\left((A_r)_{ij} - (D_r)_{ij}\right) v^{(i)} = \sum_{i=1}^{n}\left(A_{i,r:j} - D_{i,r:j}\right) v^{(i)} \\
&= V\left(A_{\cdot,r:j} - D_{\cdot,r:j}\right) \in \mathbb{R}^{1+L_{\mathrm{tot}}+q},
\end{aligned}$$

where the third equality follows from

$$\begin{aligned}
\nabla_{\theta_{rj}} \log\left\{\sum_{l=1}^{L_r}\exp(\theta_{rl}^\top v^{(i)})\right\} &= \frac{\sum_{l=1}^{L_r}\left[\exp(\theta_{rl}^\top v^{(i)})\mathbb{1}[l = j]v^{(i)}\right]}{\sum_{l=1}^{L_r}\exp(\theta_{rl}^\top v^{(i)})} \\
&= \sum_{l=1}^{L_r}\mathbb{1}[l = j]\frac{\exp(\theta_{rl}^\top v^{(i)})}{\sum_{l=1}^{L_r}\exp(\theta_{rl}^\top)} v^{(i)} \\
&= p(x_r = j | v^{(i)}) v^{(i)}
\end{aligned}$$

and $\nabla_{\theta_{rj}} \theta_{rk}^\top v^{(i)} = \mathbb{1}[k=j] v^{(i)}$. For the second-to-last line we used the definition

$$A_r = \frac{\exp(W_r)}{\exp(W_r)\mathbb{1}_{L_r}\mathbb{1}_{L_r}^T}, \qquad A = \left(A_1 \mid A_2 \mid \cdots \mid A_d\right)$$

and for the last equality we stacked the $v^{(i)}$ column-wise into the matrix

$$V = \left(v^{(1)} \mid v^{(2)} \mid \cdots \mid v^{(n)}\right) \in \mathbb{R}^{(1+L_{\mathrm{tot}}+q)\times n}.$$

The gradient w.r.t. the different θ_{rj} can be stacked together such that the j-th column of

$$V\left(A_r - D_r\right) \in \mathbb{R}^{(1+L_{\mathrm{tot}}+q)\times L_r},$$

is the gradient $\nabla_{\theta_{rj}} l_r$. The first row of this gradient corresponds to the parameters $u_{r:\cdot}$, the next block of rows corresponds to the parameters $Q_{\cdot,r:\cdot}$ and the last block of rows corresponds to the parameters $R_{\cdot,r:\cdot}$. Moreover, the gradient is the r-th vertical slice of $V\left(A - D\right) \in \mathbb{R}^{(1+L_{\mathrm{tot}}+q)\times L_{\mathrm{tot}}}$. The parameters in u and R appear in only one l_r, respectively. Therefore we have the following derivatives of $l_D = \sum_r l_r$ w.r.t. them

$$\nabla_{u^\top} l_D = \mathbb{1}_n^\top\left(A - D\right) \in \mathbb{R}^{L_{\mathrm{tot}}}, \quad \text{and}$$
$$\nabla_R l_D = Y^\top\left(A - D\right) \in \mathbb{R}^{q\times L_{\mathrm{tot}}}.$$

The situation for the parameters in Q is somewhat more complicated, since each discrete interaction parameter appears in *two* discrete node conditionals. However, we do not explicitly calculate the derivatives of l_r w.r.t. other vectors $\theta_{\hat{r}k}$ with $r \neq \hat{r}$. This is because all non-zero derivatives that appear in $\nabla_{\theta_{\hat{r}k}} l_r$ are w.r.t. parameters that are already contained in some θ_{rl}. By the symmetry $q_{rr'}(k,l) = q_{r'r}(l,k)$ these derivatives coincide. We can account for their effect by adding the transpose, that is, letting $\hat{\Phi} = \left\{\nabla_{Q_{\cdot,r:\cdot}} l_r\right\}_{r\in[d]} = D^\top\left(A - D\right) \in \mathbb{R}^{L_{\mathrm{tot}}\times L_{\mathrm{tot}}}$ we have

$$\nabla_Q l_D = \hat{\Phi} + \hat{\Phi}^\top - \mathrm{diag}_{\mathcal{B}}\left(\hat{\Phi} + \hat{\Phi}^\top\right) \in \mathbb{R}^{L_{\mathrm{tot}}\times L_{\mathrm{tot}}},$$

where we also set the *block diagonal* to zero (since the parameters on the diagonal of Q have no meaning as we store the univariate parameters separately).

If Q has been formed via $Q_u + Q_u^\top$ with an upper-triangular matrix Q_u, then the gradient stays

$$\nabla_{Q_u} l_D = \mathrm{triu}\left(\hat{\Phi} + \hat{\Phi}^\top - \mathrm{diag}_{\mathcal{B}}\left(\hat{\Phi} + \hat{\Phi}^\top\right)\right),$$

where triu set all entries that do not belong to the upper triangle to zero.

A.3 Gaussian node conditional

Let

$$\begin{aligned}\mu_s = \mu_s(x, y_{-s}) &= \alpha_s + \sum_r \rho_{sr}(x_r) - \sum_{s\neq t} \beta_{st} y_t \\ &= \alpha_s + \sum_r \rho_{sr}(x_r) - \frac{1}{2}\sum_{s\neq t}\left(\beta_{st} + \beta_{ts}\right) y_t.\end{aligned}$$

A Gaussian node conditional is described by

$$-\log p(y_s|x, y_{-s}) = \frac{1}{2}\log(2\pi) - \frac{1}{2}\log\beta_{ss} + \frac{\beta_{ss}}{2}\left(y_s - \frac{\mu_s}{\beta_{ss}}\right)^2. \tag{17}$$

In the following, we leave out the constant part for minimization.

Now we use data and define (with overloading notation)

$$\mu_s = \alpha_s \mathbb{1}_n + DR_{s,\cdot}^\top - YB_{\cdot,s} = \alpha_s\mathbb{1}_n + DR_{s,\cdot}^\top - \frac{1}{2}Y\left(B_{s,\cdot}^\top + B_{\cdot,s}\right) \in \mathbb{R}^n.$$

The square loss $l_s = -\sum_{i=1}^n \log p(y_s^{(i)}|x^{(i)}, y_{-s}^{(i)})$ for the s-th continuous variable is (up to constants) given by

$$\begin{aligned} l_s &= -\frac{n}{2}\log\beta_{ss} + \frac{\beta_{ss}}{2}\left\| \frac{\mu_s}{\beta_{ss}} - y_s \right\|_2^2 \\ &= -\frac{n}{2}\log\beta_{ss} + \frac{1}{2}\left\| \left(\frac{\mu_s}{\beta_{ss}} - y_s\right)\sqrt{\beta_{ss}} \right\|_2^2 \end{aligned}$$

Let $M = \mathbb{1}_n\alpha^\top + DR^\top - YB \in \mathbb{R}^{n\times q}$ be the vertical concatenation of the μ_s. The sum over the square loss of *all* continuous variables $l_G = \sum_s l_s$ can then be written more succinctly as

$$l_G = -\frac{n}{2}\sum_s \log\beta_{ss} + \frac{1}{2}\left\| \left(M \operatorname{diag}\left(\frac{1}{\beta}\right) - Y\right)\operatorname{diag}\left(\sqrt{\beta}\right) \right\|_F^2$$

Gradients. Let the residual $\Delta := M\operatorname{diag}\left(\frac{1}{\beta}\right) - Y \in \mathbb{R}^{n\times q}$. Analogously to the discrete node conditionals, the interaction parameters $\beta_{st}, s \neq t$ between Gaussians appear in the s-th and t-th node conditionals, respectively. We have $\Delta_{\cdot,s} = \frac{\mu_s}{\beta_{ss}} - y_s$ and recall that $\mu_s = \alpha_s\mathbb{1}_n + DR_{s,\cdot}^\top - YB_{\cdot,s}$ and $l_s = -\frac{n}{2}\log\beta_{ss} + \frac{\beta_{ss}}{2}\left\|\frac{\mu_s}{\beta_{ss}} - y_s\right\|_2^2$. By chain rule it holds

$$\partial_{\beta_{ts}} l_s = \partial_{\beta_{st}} l_s = \frac{\beta_{ss}}{2} 2\Delta_{\cdot,s}\frac{-y_t^\top}{2\beta_{ss}} = -\frac{1}{2}(Y^\top\Delta)_{st}$$

since $\partial_{\beta_{st}}\mu_s = \partial_{\beta_{ts}}\mu_s = -\frac{1}{2}y_t$. Similarly, we get $\partial_{\beta_{ts}} l_t = \partial_{\beta_{st}} l_t = -\frac{1}{2}(Y^\top\Delta)_{ts}$. This yields the following overall derivative

$$\nabla_B l_G = -\frac{1}{2}\left(Y^\top\Delta + \Delta^\top Y\right) + \frac{1}{2}\operatorname{diag}\left(Y^\top\Delta + \Delta^\top Y\right) \in \mathbb{R}^{q\times q}.$$

The other derivates are simpler and can be obtained similarly:

$$\begin{aligned} \nabla_{\alpha^\top} l_G &= \mathbb{1}_n^\top\Delta \\ \nabla_{R^\top} l_G &= D^\top\Delta \\ \nabla_{\beta_{ss}} l_G = \nabla_{\beta_{ss}} l_s &= -\frac{n}{2\beta_{ss}} + \frac{1}{2}\left\|\Delta_{:,s}\right\|_2^2 - \frac{1}{\beta_{ss}}\Delta_{:,s}^\top\mu_s, \end{aligned}$$

where the multiplication rule was used for the last gradient.

Now, suppose that $\Lambda = B + \operatorname{diag}(\beta)$ is actually parametrized via $\Lambda = FF^\top$ (a representation that is useful to enforce the PD constraint). Then, by chain rule

$$\nabla_F l_G = 2\nabla_\Lambda l_G \cdot F = 2\left(\nabla_B l_G + \operatorname{diag}(\nabla_\beta l_G)\right)\cdot F.$$

It is important not to initialize F at zero since otherwise this gradient is always zero. Indeed, the objective in F is actually non-convex with a local maximum at zero. Starting anywhere else however should lead the solver away from this local maximum.

A.4 Summary of node conditionals and gradient

Here we summarize the gradients for $l = \frac{1}{n}(l_G + l_D)$.

$$
\begin{aligned}
\nabla_B \, l &= -\frac{1}{2n}\left(Y^\top \Delta + \Delta^\top Y - \operatorname{diag}\left(Y^\top \Delta + \Delta^\top Y\right)\right)\\
\nabla_{\alpha^\top} \, l &= \frac{1}{n}\mathbb{1}_n^\top \Delta\\
\nabla_{\beta_{ss}} \, l &= -\frac{1}{2\beta_{ss}} + \frac{1}{2n}\left\|\Delta_{\cdot,s}\right\|_2^2 - \frac{1}{\beta_{ss} n}\Delta_{\cdot,s}^\top \mu_s \quad \text{with} \quad \mu_s = \alpha_s \mathbb{1}_n + D R_{s,\cdot}^\top + Y B_{\cdot,s}\\
\nabla_{u^\top} \, l &= \frac{1}{n}\mathbb{1}_n^\top \left(A - D\right) \in \mathbb{R}^{L_{\mathrm{tot}}}\\
\nabla_R \, l &= \frac{1}{n} Y^\top \left(A - D\right) + \frac{1}{n}\Delta^\top D \in \mathbb{R}^{q \times L_{\mathrm{tot}}}\\
\nabla_Q \, l &= \frac{1}{n}\left(\hat{\Phi} + \hat{\Phi}^\top - \operatorname{diag}_{\mathcal{B}}\left(\hat{\Phi} + \hat{\Phi}^\top\right)\right), \quad \text{with} \quad \hat{\Phi} = D^\top \left(A - D\right).
\end{aligned}
$$

s