COVARIANCE ESTIMATION WITH PIECEWISE-SMOOTH REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Covariance estimation for ordered variables is often built on a global regularity assumption, yet many covariance surfaces are only locally smooth: dependence evolves gradually within regimes and changes slope at a small number of locations. This paper addresses the estimation of such piecewise-smooth covariance matrices under a positive semidefinite constraint. We formulate this structure as sparsity of masked second divided differences, where the mask leaves the central variance coordinate of each stencil unpenalized, preserving near-diagonal dependence while allowing marginal variances to vary freely. The estimator combines a Frobenius data-fidelity term with a folded-concave penalty, and we develop an efficient ADMM-based algorithm that provably converges to a first-order stationary point. Under sub-Gaussian sampling, we further establish a finite-sample oracle inequality showing that the error decomposes into the oracle approximation error and a selection cost of order $\log p/n$ per unknown slope change. Simulations and an electricity-load forecasting task confirm that the method captures local slope changes that global smoothness models miss.

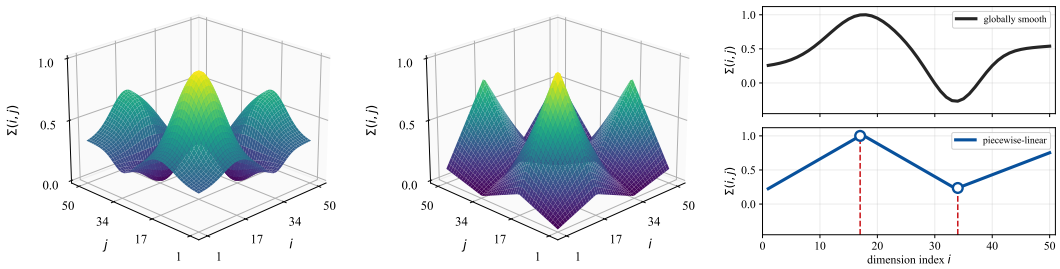
1 INTRODUCTION

Covariance matrix estimation is a fundamental problem in multivariate data analysis. The covariance matrix characterizes the second-order dependence among variables and is widely used in applications such as principal component analysis (Abdi & Williams, 2010), discriminant analysis (Zhao et al., 2024), and portfolio optimization (Ban et al., 2018). The sample covariance matrix can be noisy in finite samples and becomes singular when the dimension p exceeds sample size n. To obtain more accurate covariance matrix estimates, structural regularization is typically imposed. Representative approaches include inducing element-wise sparsity through thresholding (Bickel & Levina, 2008; Rothman et al., 2009), adopting banding and tapering methods,(Cai et al., 2010), and shrinking the covariance matrix toward a simple and well-conditioned target matrix (Ledoit & Wolf, 2004). These approaches exploit, sparsity, global distance decay, and target shrinkage, respectively.

We instead consider ordered variables whose covariance structure is piecewise-smooth along the coordinate axis. Dependence evolves gradually within local regimes but may change at a small number of unknown boundaries, rather than following a single global distance-decay pattern. This structure is relevant, for example, to yield curves across maturities, where dependence dynamics can differ across maturity ranges and market regimes (Bansal & Zhou, 2002; Gong & Remolona, 1997; Horváth et al., 2022), and to daily electricity-load curves across time, where load variability and dependence structures vary over the course of the day, particularly around periods of rising demand and peak consumption (Cottet & Smith, 2003; Cho et al., 2013; Franco et al., 2023).

We model this behavior through a piecewise-linear covariance surface: most local covariance profiles are affine along the ordered coordinates, while a sparse set of locations changes their local slopes. Figure 1 illustrates this structure with case in $p = 50$. The resulting complexity is neither entrywise sparsity nor a global distance-decay assumption: the covariance matrix itself may be dense, but its representation through local slope changes is sparse. The challenge is to estimate this structure while preserving covariance validity and the distinct roles of variances and cross-covariances. Our approach regularizes second divided differences of the covariance surface. Their sparsity induces piecewise-linear local profiles because they vanish within linear segments and are

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) Globally smooth (b) Piecewise smooth (c) Fix $j = 17$, move along i

Figure 1: The difference between globally smooth and piecewise-linear in the case with $p = 50$.

nonzero at slope changes. Divided differences also accommodate irregular grids by normalizing for interval lengths.

However, covariance estimation cannot be directly treated as a scalar problem or handled using trend filtering on a graph. The fundamental challenge is that the object of interest must be a symmetric positive semidefinite (PSD) matrix. On the one hand, the PSD constraint makes the optimization problem nonlinear in the parameter space. On the other hand, symmetry couples the difference operators along the row and column directions: applying a difference penalty to row i simultaneously modifies Σ_{ij}, which, due to $\Sigma_{ij} = \Sigma_{ji}$, directly perturbs the structure of column j. Consequently, smoothing along the rows and columns cannot be decoupled as it can for independent vectors.

To address this issue, we construct a masked second-order difference operatoron the space of symmetric matrices. Specifically, for each three-point stencil, the mask excludes the variance term on the central diagonal from penalization. This design serves two purposes. First, it preserves the near-diagonal dependence information needed to characterize the covariance surface. Second, it provides sufficient flexibility for the marginal variances, allowing them to adapt freely to their own heteroscedastic or nonsmooth variations.

Building on the above ideas, We combine a Frobenius data-fidelity term, the masked operator, and the folded-concave penalty. The penalty suppresses small slope changes while reducing bias on large ones. We further impose a positive definiteness constraint on the covariance matrix to ensure that the estimate is invertible, thereby extending the applicability of our method to downstream tasks that require the inverse covariance matrix.

In summary, our main contributions are as follows:

- We propose a piecewise-smooth covariance estimator based on a masked divided-difference operator and a folded-concave penalty. The mask leaves central variance coordinates unpenalized while retaining near-diagonal dependence information. The nonconvex regularizer preserves large structural changes and avoids excessive shrinkage.
- We develop an efficient and direct ADMM-based algorithm for solving the resulting nonconvex optimization problem and establish its convergence to a first-order stationary point.
- Under mild assumptions, we establish finite-sample error bounds with a support-dependent calibration that separates oracle estimation error from from the selection error associated with changes in local slopes.

2 Related Work

Smooth Covariance Estimation. Smooth covariance estimation typically uses trajectory smoothing, surface smoothing, low-rank eigenfunction models, or smooth Cholesky factors (Besse & Ramsay, 1986; Ramsay & Dalzell, 1991; Ramsay & Silverman, 2005; Hall et al., 2006; Staniswalis & Lee, 1998; Yao et al., 2005; Xiao et al., 2013; 2016; 2018; Huang et al., 2008; Pourahmadi, 1999; Ramsay, 2017; Mbaka et al., 2026). These methods usually assume global smoothness or a low-

dimensional smooth representation. We instead model sparse local slope changes directly in a PSD covariance matrix for ordered variables.

Trend Filtering and Adaptive Difference Regularization. Trend filtering estimates piecewise-smooth signals through sparse discrete differences (Tibshirani et al., 2005; Kim et al., 2009; Tibshirani, 2014), and graph trend filtering extends this idea to graph signals (Wang et al., 2016). Folded-concave penalties reduce the bias of convex sparsity penalties (Fan & Li, 2001; Zhang, 2010; Varma et al., 2020). Our method adapts these ideas to covariance matrices through a masked operator, PSD optimization, and sample-covariance noise.

3 Problem Formulation

In this section, we first introduce the ordered covariance surface and second divided differences, then define a masked second-order difference operator tailored to symmetric matrices, and finally formulate the optimization problem with folded-concave penalties. Correspondingly, the optimization algorithm is presented in Section 4, and the statistical theory is developed in Section 5.

3.1 Data, Ordered Coordinates, and the Covariance Surface

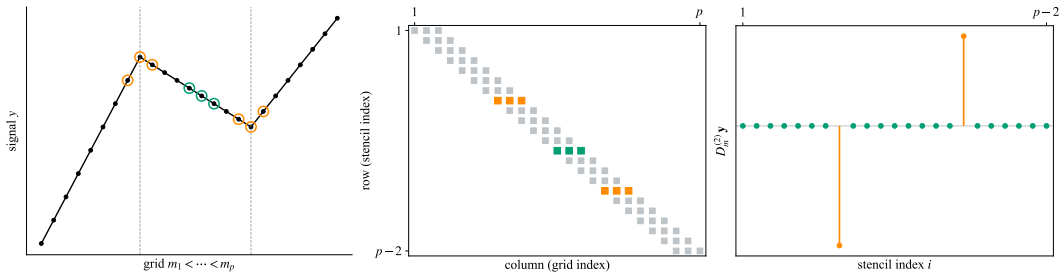
Let $\boldsymbol{x} \in \mathbb{R}^p$ be a zero-mean random vector with covariance matrix $\boldsymbol{\Sigma}^\star = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$. We observe n independent and identically distributed samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$, from which we can construct the sample covariance matrix $\boldsymbol{S} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i\boldsymbol{x}_i^\top$. Each of the p variables is associated with a point on an ordered grid $m_1 < m_2 < \cdots < m_p$, where the grid records the measurement coordinate of each variable, such as maturity or sampling time. The grid need not be equally spaced. Thus, the target covariance matrix $\boldsymbol{\Sigma} \in \mathbb{S}(p)$, where $\mathbb{S}(p)$ denotes the space of $p \times p$ symmetric matrices, can be viewed as a surface defined on the two-dimensional grid $\{(m_i, m_j)\}_{i,j=1}^p$. We assume that, along each row (equivalently, each column), this surface is piecewise-smooth, with slope changes occurring at only a small number of locations.

3.2 Divided differences and piecewise-linear profiles

To make the piecewise-smooth structure concrete, we consider the piecewise-linear case, which is naturally characterized by a second-order difference operator. Let $\boldsymbol{D} = \boldsymbol{D}_m^{(2)} \in \mathbb{R}^{(p-2)\times p}$ denote the second divided-difference matrix on the ordered grid $m_1 < \cdots < m_p$. For $i = 1, \ldots, p-2$, the i-th row of $\boldsymbol{D}$ acts on the three-point stencil $\{m_i, m_{i+1}, m_{i+2}\}$. For $k = 0, 1, 2$, we define the nonzero entries as

$$\boldsymbol{D}_{i,i+k} = \frac{1}{\prod_{r=0,\, r\neq k}^{2}(m_{i+k} - m_{i+r})}.$$

Then, $(\boldsymbol{D}\boldsymbol{y})_i = [m_i, m_{i+1}, m_{i+2}]\boldsymbol{y}$ is the second divided difference of y at m_i, m_{i+1}, m_{i+2}. Equivalently, it equals the coefficient of the quadratic term in the unique polynomial interpolating the three points $\{(m_{i+k}, y_{i+k})\}_{k=0}^2$. Thus, $(\boldsymbol{D}\boldsymbol{y})_i = 0$ if and only if these three points are collinear. Consequently, sparsity in $\boldsymbol{D}\boldsymbol{y}$ implies a piecewise-linear structure, with nonzero entries marking locations where the local slope changes.



(a) Piecewise linear signal (b) The difference matrix $\boldsymbol{D}$ (c) Sparsity of $\boldsymbol{D}\boldsymbol{y}$.

Figure 2: Second-order divided differences on an ordered grid

Figure 2 illustrates this mechanism on a single row of the covariance matrix. Panel a shows that stencils entirely within a linear segment (cyan) are collinear, yielding zero differences. Conversely, stencils crossing a knot (orange) are non-collinear and yield nonzero differences. Panels b and c confirm that the nonzero entries of $\boldsymbol{Dy}$ correspond exactly to the highlighted rows of $\boldsymbol{D}$ crossing the knots. Moreover, the sign of each nonzero difference indicates the direction of the slope change: a positive value corresponds to an upward kink (slope increases), while a negative value corresponds to a downward kink (slope decreases). Thus, the nonzero coordinates of $\boldsymbol{Dy}$ pinpoint the exact local slope changes targeted by our regularization, whereas the unpenalized zero coordinates correspond to linear segments. Moving from a single row to a full covariance matrix, symmetry couples the difference operators along the row and column directions. To account for this, we introduce a masked operator in Section 3.3. Importantly, the core mechanism remains unchanged: nonzero differences continue to identify local slope changes.

3.3 Masked Difference Operator on Covariance Matrices

As discussed above, symmetry couples the difference operators along the row and column directions, preventing them from being treated independently. This coupling can be seen explicitly at the entry level. The matrix $\boldsymbol{D\Sigma}$ collects the divided differences of $\boldsymbol{\Sigma}$ along the rows, while its transposed counterpart $\boldsymbol{\Sigma D}^\top$ collects the divided differences along the columns. Since $\boldsymbol{\Sigma}$ is symmetric, these two collections are identical entrywise: for any row index i and column index j,

$$(\boldsymbol{\Sigma D}^\top)_{j,i} = \sum_{\ell=1}^{p} \boldsymbol{\Sigma}_{j\ell}\boldsymbol{D}_{i,\ell} = \sum_{\ell=1}^{p} \boldsymbol{D}_{i,\ell}\boldsymbol{\Sigma}_{\ell j} = (\boldsymbol{D\Sigma})_{i,j},$$

where the second equality follows from $\boldsymbol{\Sigma}_{j\ell} = \boldsymbol{\Sigma}_{\ell j}$. In other words, the row-wise and column-wise divided differences are not two independent collections of linear forms, but rather different arrangements of the same collection. Penalizing them separately would therefore result in double penalization of the same linear forms.

To emphasize the center of each stencil, we re-index the rows of $\boldsymbol{D}$ by their centers: row i is now indexed by $c = i + 1$. We index each slope-change coordinate only once, using the center of its corresponding stencil as the index.

Motivated by the structural assumption that the dependence profiles, rather than the marginal scales, are piecewise-linear, we exclude the center coordinate $j = c$, whose stencil contains the central variance $\boldsymbol{\Sigma}_{cc}$. Specifically, define the index set

$$\mathcal{P} = \{(c, j) : 2 \le c \le p - 1,\ 1 \le j \le p,\ j \ne c\},$$

and define a single linear mapping $\boldsymbol{L} : \mathbb{S}(p) \to \mathbb{R}^{\mathcal{P}}$ by

$$(\boldsymbol{L\Sigma})_{(c,j)} = (\boldsymbol{D\Sigma})_{c-1,j}, \quad (c, j) \in \mathcal{P}.$$

For a fixed center c, $\boldsymbol{L}$ retains the edge coordinates $(c, c \pm 1)$ and excludes (c, c): the mask leaves marginal variances unpenalized, while diagonal entries at stencil boundaries remain regularized, preserving smoothing of the near-diagonal dependence structure.

3.4 Proposed Estimator

We formalize the piecewise-linear structural assumption as sparsity of $\boldsymbol{L\Sigma}^\star$. This assumption is structural rather than technical. It encodes the belief that the covariance surface is piecewise-linear with only a few kinks. The assumption is needed because the regularization targets precisely these slope changes. If the true surface had many kinks, no sparse penalty could recover them consistently.

Assumption 3.1. The true covariance matrix $\boldsymbol{\Sigma}^\star$ satisfies $\|\boldsymbol{L\Sigma}^\star\|_0 \le s$ for some integer s with $s \ll |\mathcal{P}|$.

Thus, the covariance profiles over most three-point stencils are locally affine, and the nonzero coordinates of $\boldsymbol{L\Sigma}^\star$ identify the locations of slope changes.

To exploit the sparsity of $\boldsymbol{L\Sigma}^\star$, We regularize $\boldsymbol{L\Sigma}$ with a folded-concave penalty to reduce shrinkage bias on genuine large slope changes. Given a penalty function $\rho_\lambda : [0, \infty) \to [0, \infty)$, define

$$P_\lambda(\boldsymbol{z}) = \sum_{q \in \mathcal{P}} \rho_\lambda(|z_q|), \qquad \boldsymbol{z} \in \mathbb{R}^{\mathcal{P}}.$$

Combining the above terms, we propose the following estimator:

$$\widehat{\boldsymbol{\Sigma}} \in \arg\min_{\boldsymbol{\Sigma}\in\mathbb{S}_+(p)} \frac{1}{2}\|\boldsymbol{\Sigma}-\boldsymbol{S}\|_{\mathrm{F}}^2 + P_\lambda(\boldsymbol{L}\boldsymbol{\Sigma}). \tag{1}$$

The estimator regularizes local slope changes directly and applies weaker shrinkage to large changes than an ℓ_1 penalty.

To facilitate the theoretical analysis, we impose the following mild conditions on the penalty function. These conditions are satisfied by common folded-concave penalties such as smoothly clipped absolute deviation (SCAD) Fan & Li (2001) and minimax concave penalty (MCP) Zhang (2010), which both have the closed-form scalar proximal maps. In our paper, we use the MCP.

Assumption 3.2 (Folded-concave penalty)**.** ρ_λ is non-decreasing with $\rho_\lambda(0)=0$, and satisfies

(i) ρ_λ is differentiable and concave on $(0,\infty)$, and its right derivative at the origin equals λ;

(ii) there exists a constant $\mu \geq 0$ such that the function $t \mapsto \rho_\lambda(t) + \frac{\mu}{2}t^2$ is convex on $[0,\infty)$.

This assumption is standard in the field related to nonconvex penatlies Fan et al. (2018); Wei & Zhao (2023). The modulus μ quantifies the weak concavity of the nonconvex penalty and is the basic quantify in the algorithm convergence and statistical analyses below.

4 ADMM-BASED ALGORITHM

To solve Problem 1, we propose an efficient alternating direction method of multipliers (ADMM). We start with introducing the auxiliary variables $\boldsymbol{Z}=\boldsymbol{L}\boldsymbol{\Sigma}, \boldsymbol{V}=\boldsymbol{\Sigma}$. Then Problem 1 is equivalent to the following form:

$$\begin{aligned} \min_{\boldsymbol{\Sigma},\boldsymbol{Z},\boldsymbol{V}} \quad & \frac{1}{2}\|\boldsymbol{\Sigma}-\boldsymbol{S}\|_{\mathrm{F}}^2 + P_\lambda(\boldsymbol{Z}) + \iota_{\mathbb{S}_+(p)}(\boldsymbol{V}) \\ \text{s.t.} \quad & \boldsymbol{Z}=\boldsymbol{L}\boldsymbol{\Sigma}, \qquad \boldsymbol{V}=\boldsymbol{\Sigma}, \end{aligned} \tag{2}$$

where $\iota_{\mathbb{S}_+(p)}(\cdot)$ denotes the indicator function of the positive semidefinite cone $\mathbb{S}_+(p)$, defined as $\iota_{\mathbb{S}_+(p)}(\boldsymbol{X})=0$ if $\boldsymbol{X}\in\mathbb{S}_+(p)$ and $+\infty$ otherwise. For applications requiring an invertible estimate, we use the strict-PD variant obtained by replacing $\mathbb{S}_+(p)$ with

$$\mathbb{S}_{+,\delta}(p) = \{\boldsymbol{\Sigma}\in\mathbb{S}(p) : \boldsymbol{\Sigma}\succeq\delta I\}, \qquad \delta\geq 0,$$

where choice $\delta=0$ recovers Problems 1.

Let $\boldsymbol{U},\boldsymbol{A}$ be the scaled dual variables associated with the two constraints. The scaled augmented Lagrangian of Problem 2 is

$$\begin{aligned} \mathcal{L}_\tau(\boldsymbol{\Sigma},\boldsymbol{Z},\boldsymbol{V},\boldsymbol{U},\boldsymbol{A}) = & \frac{1}{2}\|\boldsymbol{\Sigma}-\boldsymbol{S}\|_F^2 + P_\lambda(\boldsymbol{Z}) + \iota_{\mathbb{S}_{+,\delta}(p)}(\boldsymbol{V}) \\ & + \frac{\tau}{2}\left(\|\boldsymbol{L}\boldsymbol{\Sigma}-\boldsymbol{Z}+\boldsymbol{U}\|_F^2 + \|\boldsymbol{\Sigma}-\boldsymbol{V}+\boldsymbol{A}\|_F^2 - \|\boldsymbol{U}\|_F^2 - \|\boldsymbol{A}\|_F^2\right). \end{aligned} \tag{3}$$

Since the last two terms are independent of $\boldsymbol{\Sigma},\boldsymbol{Z},\boldsymbol{V}$, they can be omitted when deriving the primal updates. In the $(t+1)$-th iteration, our proposed algorithm has the following updating steps:

$$\boldsymbol{\Sigma}^{t+1} = \arg\min_{\boldsymbol{\Sigma}} \left\{ \frac{1}{2}\|\boldsymbol{\Sigma}-\boldsymbol{S}\|_{\mathrm{F}}^2 + \frac{\tau}{2}\left\|\boldsymbol{L}\boldsymbol{\Sigma}-\boldsymbol{Z}^t+\boldsymbol{U}^t\right\|_{\mathrm{F}}^2 + \frac{\tau}{2}\left\|\boldsymbol{\Sigma}-\boldsymbol{V}^t+\boldsymbol{A}^t\right\|_{\mathrm{F}}^2 \right\}, \tag{4}$$

$$\boldsymbol{Z}^{t+1} = \mathrm{prox}_{P_\lambda/\tau}\left(\boldsymbol{L}\boldsymbol{\Sigma}^{t+1}+\boldsymbol{U}^t\right), \quad \boldsymbol{V}^{t+1} = \mathrm{Proj}_{\mathbb{S}_{+,\delta}(p)}\left(\boldsymbol{\Sigma}^{t+1}+\boldsymbol{A}^t\right). \tag{5}$$

followed by the dual updates $\boldsymbol{U}^{t+1}=\boldsymbol{U}^t+\boldsymbol{L}\boldsymbol{\Sigma}^{t+1}-\boldsymbol{Z}^{t+1}$ and $\boldsymbol{A}^{t+1}=\boldsymbol{A}^t+\boldsymbol{\Sigma}^{t+1}-\boldsymbol{V}^{t+1}$.

The $\boldsymbol{\Sigma}$-update is a strongly convex quadratic with unique minimizer

$$\left((1+\tau)I+\tau\boldsymbol{L}^\top\boldsymbol{L}\right)\boldsymbol{\Sigma}^{t+1} = \boldsymbol{S}+\tau\boldsymbol{L}^\top(\boldsymbol{Z}^t-\boldsymbol{U}^t)+\tau(\boldsymbol{V}^t-\boldsymbol{A}^t), \tag{6}$$

which can be solved by efficient algorithms, such as conjugate gradients.

The $\boldsymbol{Z}$-update is coordinatewise and has a closed form for MCP. Because P_λ is separable, it reduces to a series of independent scalar problems

$$\min_{z\in\mathbb{R}} \ \rho_\lambda(|z|) + \frac{\tau}{2}(z-v)^2, \qquad v \in \mathbb{R},$$

whose global minimizer can obtained by proximal operator Huang et al. (2012). Moreover, under Assumption 3.2, we require $\tau > \mu$, which ensures that the scalar objective is convex and hence admits a unique minimizer.

The $\boldsymbol{V}$-update is an eigenvalue projection. Let

$$\boldsymbol{\Sigma}^{t+1} + \boldsymbol{A}^t = \boldsymbol{Q}\,\mathrm{diag}(d_1,\dots,d_p)\boldsymbol{Q}^\top$$

be an eigen-decomposition, where $\mathrm{diag}(d_1,\dots,d_p)$ denotes the diagonal matrix with diagonal entries $d_1,\dots,d_p$. Then one has

$$\boldsymbol{V}^{t+1} = \boldsymbol{Q}\,\mathrm{diag}\big(\max\{d_1,\delta\},\dots,\max\{d_p,\delta\}\big)\boldsymbol{Q}^\top.$$

Thus the update clips eigenvalues below δ, and reduces to the usual PSD projection when $\delta = 0$.

Despite the presence of three primal variables, the joint $(\boldsymbol{Z},\boldsymbol{V})$-update given $\boldsymbol{\Sigma}^{t+1}$ places the iteration within the standard two-block ADMM framework Boyd et al. (2011), which enables the convergence analysis in Appendix B.

Theorem 4.1 (Convergence to a stationary point)**.** *Let Assumption 3.2 hold, set $\kappa_K = \|\boldsymbol{L}\|_{\mathrm{op}}^2 + 1$, where $\|\boldsymbol{L}\|_{\mathrm{op}}$ denotes its operator norm, and suppose that a split KKT point of Problem 1 exists. If there is an η such that*

$$0 < \eta < 1 - \mu\kappa_K, \qquad \tau > 2\mu + \frac{\mu^2\kappa_K}{\eta},$$

then the proposed ADMM-based algorithm converges to a stationary point of Problem 1.

These are sufficient conditions for the exact, fixed-parameter iteration. In particular, they imply $\tau > \mu$, so each $\boldsymbol{Z}$-subproblem is uniquely defined.

5 Statistical Properties

In this section, we establish the statistical guarantee for our proposed estimator. It is formulated through the first-order stationary condition. Its complexity is governed by the true slope-change support, the oracle subspace induced by that support, and a support-dependent Green scale that calibrates the tuning parameter.

5.1 Stationarity, support, and the Green scale

Write $\boldsymbol{\varepsilon} = \boldsymbol{S} - \boldsymbol{\Sigma}^\star$ for the sample-covariance noise, and denote by $\Pi_{\mathcal{U}}$ the orthogonal projection onto a subspace $\mathcal{U} \subseteq \mathbb{S}(p)$. Our analysis rests on three quantities: the stationarity condition that identifies the estimator, the support of the true slope changes, and the noise scale that calibrates the tuning parameter.

A feasible matrix $\widehat{\boldsymbol{\Sigma}}$ is a first-order stationary point (FOSP) of Problem 1 if

$$\mathbf{0} \in \widehat{\boldsymbol{\Sigma}} - \boldsymbol{S} + \boldsymbol{L}^\top \partial P_\lambda(\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}) + \mathcal{N}_{\mathbb{S}_+(p)}(\widehat{\boldsymbol{\Sigma}}), \tag{7}$$

where $\mathcal{N}_{\mathbb{S}_+(p)}$ is the normal cone of the PSD cone and ∂ denotes the limiting subdifferential. Since the loss is quadratic and the penalty is weakly convex, the FOSP condition identifies the estimator in the regime where the curvature of the loss dominates the concavity of the penalty, and the minimizer is then unique.

The structural complexity of the truth is carried by the slope-change support $\mathcal{A} = \{q \in \mathcal{P} : (\boldsymbol{L}\boldsymbol{\Sigma}^\star)_q \neq 0\}$ of cardinality $K = |\mathcal{A}|$: the locations where the covariance surface departs from local affine behavior. Knowledge of $\mathcal{A}$ would reduce estimation to fitting only these K coordinates, with all remaining coordinates of $\boldsymbol{L}\boldsymbol{\Sigma}^\star$ forced to zero. The matrices satisfying these constraints form the oracle subspace $\mathcal{U}_{\mathcal{A}} = \{\boldsymbol{\Xi} \in \mathbb{S}(p) : (\boldsymbol{L}\boldsymbol{\Xi})_q = 0 \text{ for every } q \notin \mathcal{A}\}$, whose dimension

$\overline{K} = \dim \mathcal{U}_{\mathcal{A}}$ is the effective number of free parameters. Writing $\nu = \dim \ker \boldsymbol{L}$ for the unpenalized directions in which $\boldsymbol{L}$ vanishes identically, one always has $\overline{K} \le K + \nu$, the bound counting the K active coordinates plus the ν kernel directions; it may be strict when active coordinates are linearly dependent through $\boldsymbol{L}$.

Finally, the tuning parameter must dominate the dual noise on off-support coordinates. For $J \subseteq \mathcal{P}$ let $\boldsymbol{L}_J$ denote $\boldsymbol{L}$ restricted to J; the deterministic Green scale

$$g_{\mathcal{A}} = \max_{q \in \mathcal{P} \setminus \mathcal{A}} \left\| \boldsymbol{L}^{+}_{\mathcal{P} \setminus \mathcal{A}} \boldsymbol{e}_q \right\|_{\mathrm{F}}, \tag{8}$$

where $+$ denotes the Moore–Penrose pseudoinverse, depends on the grid, mask, and realized support, and measures the amplification of off-support dual noise through the masked operator.

Beyond the structural Assumption 3.1 and the penalty Assumption 3.2, we impose the following probabilistic and design conditions.

Assumption 5.1 (Sub-Gaussian observations)**.** There is $\varsigma > 0$ such that, for every unit vector $\boldsymbol{u} \in \mathbb{R}^p$ and $t \ge 0$,

$$\mathbb{P}\big(|\boldsymbol{u}^\top \boldsymbol{x}| > t\big) \le 2\exp(-t^2/\varsigma^2).$$

The Assumption 5.1 controls the tails of all one-dimensional projections of $\boldsymbol{x}$ and hence the concentration of the sample covariance around $\boldsymbol{\Sigma}^\star$, which is standard in the covariance matrix estimation Bickel & Levina (2008); Rothman et al. (2009).

Assumption 5.2 (Operator scale)**.** For the fixed grid and mask, $\|\boldsymbol{L}\|_{\mathrm{op}} \le C_L$, where C_L is a universal constant, independent of p and of the observations.

When $\mu C_L^2 < 1$, the data-fidelity term dominates the weak concavity of the penalty, so the objective in Problem 1 is strongly convex and every $\widehat{\boldsymbol{\Sigma}}$ is the unique global minimizer. The theorem below instead requires $\frac{3}{2}\mu C_L^2 < 1$, a slightly stronger condition imposed by the present oracle-inequality argument; the objective remains strongly convex in this regime.

5.2 Oracle inequality

An oracle inequality compares the estimator with the best constrained approximation of $\boldsymbol{\Sigma}^\star$: for a hypothetical set $T \subseteq \mathcal{P}$ of allowed slope-change locations, let $T^c = \mathcal{P} \setminus T$ and consider matrices that are locally affine outside T. Such a matrix pays no penalty on T^c, but each coordinate in T must be selected and controlled at the noise scale λ. We therefore measure each candidate $\boldsymbol{\Sigma}$ against

$$\mathfrak{R}_\lambda(\boldsymbol{\Sigma}, T) = \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star\|_{\mathrm{F}}^2 + 4P_{\lambda, T^c}(\boldsymbol{L}\boldsymbol{\Sigma}) + 16 C_L^2 \lambda^2 |T|,$$

where $P_{\lambda, T^c}(\boldsymbol{z}) = \sum_{q \in T^c} \rho_\lambda(|z_q|)$ is the off-set penalty on divided-difference coordinates. The three terms measure approximation error, off-support deviation from local affine behavior, and the price of the selected coordinates, respectively.

Theorem 5.3 (Oracle inequality in the curvature-dominated regime)**.** *Suppose Assumptions 3.2, 5.1, and 5.2 hold. Suppose* $\frac{3}{2}\mu C_L^2 < 1, n \ge c_0 \log p$, *let* $\zeta \ge g_{\mathcal{A}}$ *be deterministic and choose* $\lambda \ge C\zeta\sqrt{\frac{2\log p}{n}}$ *for sufficiently large constants* c_0, C *depending only on the sub-Gaussian scale. Then, with probability at least* $1 - c_2 p^{-2}$, *every* $\widehat{\boldsymbol{\Sigma}}$ *of Problem 1 satisfies*

$$\left\| \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star \right\|_{\mathrm{F}}^2 \le \frac{1}{1 - \frac{3}{2}\mu C_L^2} \inf_{\boldsymbol{\Sigma} \in \mathbb{S}_+(p),\ T \subseteq \mathcal{P}} \mathfrak{R}_\lambda(\boldsymbol{\Sigma}, T) + \frac{4\,\|\Pi_{\mathcal{U}_{\mathcal{A}}} \boldsymbol{\varepsilon}\|_{\mathrm{F}}^2}{1 - \frac{3}{2}\mu C_L^2}. \tag{9}$$

Theorem 5.3 states that the estimator adapts to the unknown support $\mathcal{A}$: it matches, up to constants, the risk of an oracle that knows $\mathcal{A}$ in advance. When $\boldsymbol{\Sigma}^\star$ is exactly piecewise-linear, the risk is driven solely by the noise projected onto $\mathcal{U}_{\mathcal{A}}$, with effective dimension $\overline{K}$ instead of the p^2 free entries of $\boldsymbol{\Sigma}^\star$.

The following corollary specializes Theorem 5.3 to the exact piecewise-linear model targeted by our estimator.

Corollary 5.4 (Rate for a piecewise-linear covariance surface)**.** *Suppose the conditions of Theorem 5.3 hold, under Assumption 3.1, with $\mathcal{A}$ and $K = |\mathcal{A}|$ as defined in Section 5.1. Then, with probability at least $1 - c_2 p^{-2}$, every $\widehat{\boldsymbol{\Sigma}}$ of Problem 1 satisfies*

$$\left\| \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star \right\|_{\mathrm{F}}^2 \leq \frac{C'}{1 - \frac{3}{2}\mu C_L^2} \frac{\overline{K} + \log p + C_L^2 \zeta^2 K \log p}{n}, \qquad \overline{K} \leq K + \nu. \tag{10}$$

Corollary 5.4 shows that with a fixed number of knots K, the estimator attains the near-parametric rate $(K + \log p)/n$, as if the support were known, while learning each unknown slope change costs only a logarithmic factor $\log p/n$.

6 Experiments

In this section, we evaluate the performance of our proposed piecewise-smooth covariance estimation method on both synthetic and real-world datasets. In addition to SCM, we primarily compare our method with smooth regularization without monotonicity method (SRCov) in Malioutov et al. (2016) and two recent ones proposed in Mbaka et al. (2026), namely, the low-rank smoothing method (LRCov) and the Cholesky smoothing method (CholCov).

6.1 Synthetic covariance estimation

We evaluate two covariance families on the uniform grid $t_i = (i-1)/(p-1)$. The matched full-rank lag family is

$$\boldsymbol{\Sigma}^\star_{ij} = f(|i-j|) + 0.15\mathbf{1}\{i = j\},$$

where f is a piecewise-linear interpolation. We also use a low-rank piecewise-linear family,

$$\boldsymbol{\Sigma}^\star = \frac{p}{\mathrm{tr}(\boldsymbol{F}\boldsymbol{F}^\top + 0.18\boldsymbol{I})}(\boldsymbol{F}\boldsymbol{F}^\top + 0.18\boldsymbol{I}),$$

where $\boldsymbol{F} \in \mathbb{R}^{p\times 4}$ has piecewise-linear loadings. These two types of $\boldsymbol{\Sigma}$ are both PSD and piecewise-smooth. Full details of above generations are given in the Appendix A. We fix the sample size $n = 100$ and run 100 Monte Carlo simulations. $p \in \{50, 100, 200\}$ is adopted for our experiments.

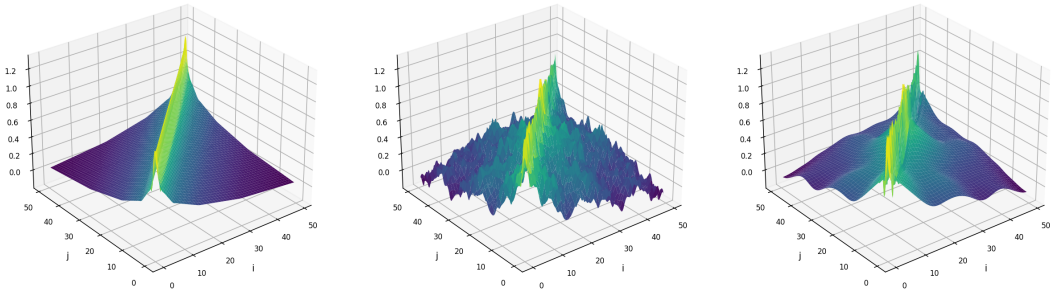
Setting $\gamma = 10$ for the MCP[1] and $\tau = 1$ for ADMM, we use five-fold cross-validation to jointly select the penalty level λ and the PSD eigenvalue floor δ, based on the Frobenius loss of the held-out covariance. We use relative Frobenius error as the main metric, which is reported in the Tabel 1. Other metrics are reported in the Appendix A.

Table 1: Mean relative Frobenius error over 100 Monte Carlo simulations with $n = 100$.

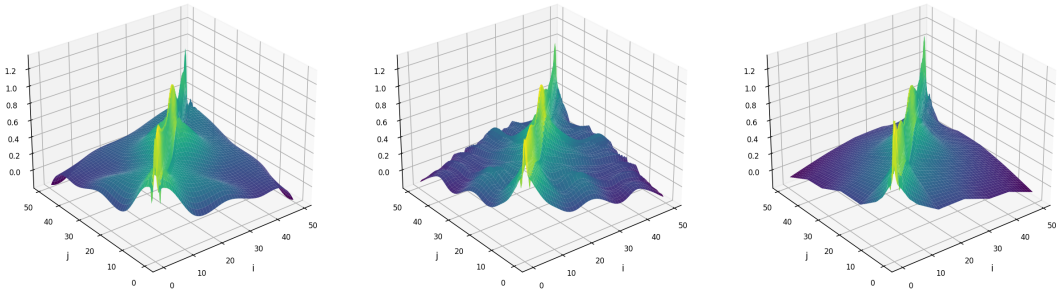
Regime	p	SCM	SRCOV	LRCov	CholCov	Ours
full-rank lag	50	0.3023	0.2773	0.2561	0.2679	**0.2128**
full-rank lag	100	0.4002	0.3250	0.3243	0.3305	**0.2665**
full-rank lag	200	0.4864	0.3832	0.4335	0.4239	**0.3261**
low-rank piecewise-linear	50	0.1576	0.2043	**0.1441**	0.2037	0.1448
low-rank piecewise-linear	100	0.1782	0.1924	0.1653	0.2534	**0.1635**
low-rank piecewise-linear	200	0.1455	0.1444	0.1292	0.1926	**0.1289**

Table 1 supports our structural motivation. Both estimators reduce the estimation error relative to the SCM. For the full-rank lag family, our proposed method outperforms LRCov and CholCov across all dimensions. It remains competitive on the low-rank family, which naturally favors low-rank covariance estimators. To further illustrate how our method captures the underlying piecewise-linear structure, Figure 3 visualizes the estimated covariance surfaces produced by the different methods for the full-rank lag family with $p = 50$, which presents that our proposed method captures the piecewise-smooth structure.

[1]The MCP is defined as $\rho_\lambda(z) = \int_0^{|z|} \left(\lambda - \frac{t}{\gamma}\right)_+ dt, \quad \gamma > 1.$



(a) Truth (b) SCM (c) SRCov



(d) LRCov (e) CholCov (f) Ours

Figure 3: Covariance surfaces for a representative full-rank lag replication with $p = 50$.

6.2 Electricity load covariance forecasting

We evaluate covariance prediction on the UCI ElectricityLoadDiagrams20112014 data. We aggregate a fixed cohort of 50 active clients into daily curves with 96 fifteen-minute coordinates. The data preprocessing is recorded in the Appendix A.3. Each observation is therefore a 96-dimensional intraday load-change curve.

Table 2: Mean held-out predictive NLL of different methods.

Metric	SCM	SRCov	LRCov	CholCov	Ours
Mean NLL	-119.13	-135.86	-131.96	-119.44	**-157.32**

We use chronological rolling windows with 730, 146, and 146 days for training, validation, and testing, respectively. After differencing, the sample sizes are 729, 145, and 145. Four test blocks are non-overlapping. The parameter λ and δ are selected on validation data. We compare the same baselines in our synthetic data experiment, using a common evaluation ridge for held-out Gaussian NLL. Because some methods produce estimates that are not positive definite, we adjust their estimates accordingly, the corresponding details are provided in the Appendix. Table 2 shows that our estimator has the best performance among all methods.

7 Conclusion

We proposed a covariance estimator for ordered variables whose dependence profile is piecewise-smooth along the coordinate axis, modeling local slope changes as a sparse masked second divided difference subject to a positive semidefinite constraint, with a folded-concave penalty that limits shrinkage bias on genuine kinks. An efficient ADMM-based algorithm provably converges to a first-order stationary point, and a finite-sample oracle inequality shows that the price of not knowing the slope-change locations is only logarithmic in the dimension. A natural next step is to extend the masked-difference framework to piecewise-polynomial covariance surfaces via higher-order difference operators.

AI USE STATEMENT

We used generative AI tools only for language polishing and editing of author-written text. AI was not used for research ideation, experiments, data analysis, code, or figures. We reviewed all AI-assisted text and take full responsibility for the final content.

ETHICS STATEMENT

The authors declare no ethical concerns. This work does not involve human subjects, sensitive data, or harmful applications. There are no conflicts of interest.

REPRODUCIBILITY STATEMENT

We provide complete experimental details, hyperparameters, data preprocessing, and evaluation protocols in the Appendix. These details are sufficient to reproduce the reported results. The code will be publicly released upon publication.

REFERENCES

Hervé Abdi and Lynne J. Williams. Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(4):433–459, 2010.

Gah-Yi Ban, Noureddine El Karoui, and Andrew E. B. Lim. Machine learning and portfolio optimization. *Management Science*, 64(3):1136–1154, 2018.

Ravi Bansal and Hao Zhou. Term structure of interest rates with regime shifts. *The Journal of Finance*, 57(5):1997–2043, 2002. doi: 10.1111/0022-1082.00487.

Philippe Besse and James O. Ramsay. Principal components analysis of sampled functions. *Psychometrika*, 51(2):285–311, 1986.

Peter J. Bickel and Elizaveta Levina. Covariance regularization by thresholding. *The Annals of Statistics*, 36(6):2577–2604, 2008.

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

T. Tony Cai, Cun-Hui Zhang, and Harrison H. Zhou. Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics*, 38(4):2118–2144, 2010.

Haeran Cho, Yannig Goude, Xavier Brossat, and Qiwei Yao. Modeling and forecasting daily electricity load curves: A hybrid approach. *Journal of the American Statistical Association*, 108(501): 7–21, 2013. doi: 10.1080/01621459.2012.722900.

Remy Cottet and Michael Stanley Smith. Bayesian modeling and forecasting of intraday electricity load. *Journal of the American Statistical Association*, 98(464):839–849, 2003. doi: 10.1198/016214503000000774.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan, Han Liu, Qiang Sun, and Tong Zhang. I-LAMM for sparse learning: Simultaneous control of algorithmic complexity and statistical error. *The Annals of Statistics*, 46(2):814–841, 2018.

Gabriel Franco, Camila P. E. de Souza, and Nancy L. Garcia. Aggregated functional data model applied on clustering and disaggregation of uk electrical load profiles. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 72(1):48–75, 2023. doi: 10.1093/jrsssc/qlac006.

Fangxiong Gong and Eli M. Remolona. Two factors along the yield curve. *The Manchester School*, 65(S):1–31, 1997. doi: 10.1111/1467-9957.65.s.1.

Peter Hall, Hans-Georg Müller, and Jane-Ling Wang. Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, 34(3):1493–1517, 2006.

Lajos Horváth, Piotr Kokoszka, Jeremy VanderDoes, and Shixuan Wang. Inference in functional factor models with applications to yield curves. *Journal of Time Series Analysis*, 43:872–894, 2022. doi: 10.1111/jtsa.12642.

Jian Huang, Patrick Breheny, and Shuangge Ma. A selective review of group selection in high-dimensional models. *Statistical Science*, 27(4):481–499, 2012.

Jianhua Z. Huang, Haipeng Shen, and Andreas Buja. Functional principal components analysis via penalized rank one approximation. *Electronic Journal of Statistics*, 2:678–695, 2008.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2):339–360, 2009.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

Dmitry M. Malioutov, Aycan A. Corum, and Mujdat Çetin. Covariance matrix estimation for interest-rate risk modeling via smooth and monotone regularization. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1006–1014, 2016.

Uchema Mbaka, James O. Ramsay, and Michelle Carey. Estimating a smooth covariance for functional data. *Computational Statistics & Data Analysis*, 213:108255, 2026.

Mohsen Pourahmadi. Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677–690, 1999.

James O. Ramsay. A functional estimate of covariation. *Journal of Computational and Graphical Statistics*, 26(1):160–170, 2017.

James O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *Journal of the Royal Statistical Society: Series B*, 53(3):539–561, 1991.

James O. Ramsay and Bernard W. Silverman. *Functional Data Analysis*. Springer, 2 edition, 2005.

Adam J. Rothman, Elizaveta Levina, and Ji Zhu. Generalized thresholding of large covariance matrices. *Journal of the American Statistical Association*, 104(485):177–186, 2009.

Joan G. Staniswalis and J. Jack Lee. Nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*, 93(444):1403–1418, 1998.

Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B*, 67(1):91–108, 2005.

Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *The Annals of Statistics*, 42(1):285–323, 2014.

Rohan Varma, Harlin Lee, Jelena Kovačević, and Yuejie Chi. Vector-valued graph trend filtering with non-convex penalties. *IEEE Transactions on Signal and Information Processing over Networks*, 6:48–62, 2020.

Yu-Xiang Wang, James Sharpnack, Alexander J. Smola, and Ryan J. Tibshirani. Trend filtering on graphs. *Journal of Machine Learning Research*, 17(105):1–41, 2016.

Quan Wei and Ziping Zhao. Large covariance matrix estimation with oracle statistical rate via majorization-minimization. *IEEE Transactions on Signal Processing*, 71:3328–3342, 2023.

Luo Xiao, Yingxing Li, and David Ruppert. Fast bivariate P-splines: The sandwich smoother. *Journal of the Royal Statistical Society: Series B*, 75(3):577–599, 2013.

Luo Xiao, Vadim Zipunnikov, David Ruppert, and Ciprian Crainiceanu. Fast covariance estimation for high-dimensional functional data. *Statistics and Computing*, 26:409–421, 2016.

Luo Xiao, Cai Li, William Checkley, and Ciprian Crainiceanu. Fast covariance estimation for sparse functional data. *Statistics and Computing*, 28(3):511–522, 2018.

Fang Yao, Hans-Georg Müller, and Jane-Ling Wang. Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100(470):577–590, 2005.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

Shuping Zhao, Bob Zhang, Jian Yang, Jianhang Zhou, and Yong Xu. Linear discriminant analysis. *Nature Reviews Methods Primers*, 4(1):70, 2024.

A Additional Experimental Results

A.1 Evaluation metrics

For a synthetic estimate $\widehat{\boldsymbol{\Sigma}}$, we report the relative Frobenius errors

$$\mathrm{RelFro}(\widehat{\boldsymbol{\Sigma}}) = \frac{\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}}{\|\boldsymbol{\Sigma}^\star\|_{\mathrm{F}}}.$$

We measure roughness relative to the truth by

$$\mathrm{Rough}(\widehat{\boldsymbol{\Sigma}}) = \frac{\mathrm{RMS}(|\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}|)}{\mathrm{RMS}(|\boldsymbol{L}\boldsymbol{\Sigma}^\star|)}, \qquad \mathrm{RMS}(|\boldsymbol{z}|) = \left(\frac{1}{|\mathcal{P}|}\sum_{q\in\mathcal{P}} z_q^2\right)^{1/2}.$$

A value near one indicates curvature magnitude comparable to the truth.

For held-out data with sample covariance $\boldsymbol{S}_{\mathrm{test}}$, we evaluate the Gaussian negative log-likelihood (NLL), up to the method-independent factor $1/2$, as

$$\mathrm{NLL}(\widehat{\boldsymbol{\Sigma}}) = \log\det(\widehat{\boldsymbol{\Sigma}} + \rho I) + \mathrm{tr}\Big\{\boldsymbol{S}_{\mathrm{test}}(\widehat{\boldsymbol{\Sigma}} + \rho I)^{-1}\Big\}.$$

Here ρ is a common evaluation ridge used for all methods. In the electricity experiment, $\boldsymbol{S}_{\mathrm{test}}$ is formed around the fixed training mean and $\rho = 10^{-4}\,\mathrm{tr}(\boldsymbol{S})/p$. Lower NLL indicates better out-of-sample covariance prediction.

A.2 Synthetic-data protocol

All $\boldsymbol{\Sigma}^\star$ are defined on the uniform grid

$$t_i = \frac{i-1}{p-1}, \qquad i = 1, \ldots, p,$$

and are trace-normalized to p. For the full-rank lag family, let $d = |i-j|$ denote the lag between coordinates i and j. We first define the two-scale exponential profile

$$g(d) = 0.55\exp(-d/20) + 0.45\exp(-d/2),$$

and let $f(d)$ be its piecewise-linear interpolation at the lag knots

$$\{0, 0.04, 0.12, 0.30, 0.60, 1\}(p-1).$$

The population covariance is then

$$\boldsymbol{\Sigma}^\star_{ij} = f(|i-j|) + 0.15\mathbf{1}\{i=j\}.$$

The diagonal addition ensures a strictly positive definite full-rank covariance while preserving the sparse local slope-change structure induced by the piecewise-linear lag profile.

For the low-rank family, the four columns of $\boldsymbol{F} \in \mathbb{R}^{p\times 4}$ are piecewise-linear loading curves with interior knots at

$$\mathrm{round}\big(0.25(p-1)\big) \quad \text{and} \quad \mathrm{round}\big(0.75(p-1)\big).$$

The first loading is the affine curve

$$f_1(t) = 0.28 + 0.6t.$$

The second and third loadings are fixed piecewise-linear curves with distinct knot values and slope changes at the two interior knots. The fourth loading is constructed by independently sampling its four knot values from $\mathrm{Unif}(0.1, 0.8)$. The exact knot values and random seeds are provided in the released experiment configuration. We then define

$$\boldsymbol{\Sigma}_0 = \boldsymbol{F}\boldsymbol{F}^\top + 0.18\boldsymbol{I}, \qquad \boldsymbol{\Sigma}^\star = \frac{p\,\boldsymbol{\Sigma}_0}{\mathrm{tr}(\boldsymbol{\Sigma}_0)}.$$

Table 3: Additional synthetic results, averaged over 100 Monte Carlo simulations. Held-out NLL and roughness are reported.

Regime	Method	Held-out NLL	Roughness
lag, $p = 50$	SCM	48.1	3.57
	LRCov	22.2	2.71
	CholCov	15.8	2.09
	SRCOV	20.8	2.98
	Ours	**15.1**	**2.08**
lag, $p = 100$	SCM	69.5	5.17
	LRCov	47.3	3.22
	SRCOV	39.9	3.27
	Ours	**18.4**	**1.22**
lag, $p = 200$	SCM	134.1	7.29
	LRCov	107.5	4.07
	SRCOV	54.7	3.31
	Ours	**32.6**	**0.81**
low-rank, $p = 100$	SCM	-20.9	5.16
	LRCov	**-93.1**	1.20
	CholCov	-69.1	1.56
	SRCOV	-72.2	1.16
	Ours	-90.1	**0.89**
low-rank, $p = 50$	SCM	-5.5	3.35
	LRCov	**-41.9**	1.09
	CholCov	-35.6	1.22
	SRCOV	-31.4	1.24
	Ours	-40.0	**0.92**
low-rank, $p = 200$	SCM	-159.2	6.83
	LRCov	-189.6	**1.05**
	CholCov	-169.2	1.25
	SRCOV	-174.5	1.14
	Ours	**-197.0**	1.11

We construct the observed samples as follows. Let $\boldsymbol{x} = \boldsymbol{z} + \boldsymbol{e}$, where $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^\star)$ and $\boldsymbol{e} \sim \mathcal{N}(\mathbf{0}, 0.2I)$. Only $\boldsymbol{x}$ is observed. As stated earlier, we fix the training sample size $n = 100$ in our synthetic data. We independently generate an additional 200 samples from the same data-generating process and use them as a held-out set; these held-out samples are not used during training.

We fix MCP $\gamma = 10$ and ADMM $\tau = 1$. Five-fold cross-validation jointly chooses $\lambda = \alpha\,\mathrm{RMS}(|\boldsymbol{L}\boldsymbol{S}|)$, $\alpha \in \{1, 2, 4, 8, 16, 32\}$, and the PSD floor $\beta\,\mathrm{tr}(\boldsymbol{S})/p$, $\beta \in \{0.05, 0.1, 0.2, 0.4\}$, using held-out covariance Frobenius loss. We use ADMM tolerance at 10^{-5}.

LRCov and CholCov use the automatic GCV/RGCV-based routines from their original paper Mbaka et al. (2026).

A.3 Data pre-processing of Electricity-load experiment

We use the UCI ElectricityLoadDiagrams20112014 data, which records fifteen-minute electricity consumption for 370 clients from 2011 to 2014. After removing incomplete days, 1460 complete daily curves remain. A fixed cohort of 50 active clients is selected from an initial historical period and aggregated at each time point, yielding one 96-dimensional curve per day.

For each rolling window, weekday–month means and coordinatewise scales are estimated from training data only. We remove the seasonal means, take consecutive-day differences, and standardize by the training scales. The resulting training, validation, and test matrices contain 729, 145, and 145 observations. The strict-PD floor and MCP penalty level are selected by validation predictive NLL. All methods use the common evaluation ridge $10^{-4}\,\mathrm{tr}(\boldsymbol{S})/p$ for test NLL.

B Proof of Theorem 4.1

Write $\boldsymbol{Y} = (\boldsymbol{Z}, \boldsymbol{V})$, with

$$\boldsymbol{K}\boldsymbol{\Sigma} = (\boldsymbol{L}\boldsymbol{\Sigma}, \boldsymbol{\Sigma}), \qquad G(\boldsymbol{Y}) = P_\lambda(\boldsymbol{Z}) + \iota_{\mathbb{S}_+(p)}(\boldsymbol{V}).$$

Then Problem 2 is $\min_{\boldsymbol{\Sigma},\boldsymbol{Y}} \ \frac{1}{2}\|\boldsymbol{\Sigma}-\boldsymbol{S}\|_{\mathrm{F}}^2 + G(\boldsymbol{Y})$ subject to $\boldsymbol{K}\boldsymbol{\Sigma}=\boldsymbol{Y}$. Its adjoint is $\boldsymbol{K}^*(\boldsymbol{z},\boldsymbol{v}) = \boldsymbol{L}^*\boldsymbol{z}+\boldsymbol{v}$, and hence

$$\boldsymbol{K}^*\boldsymbol{K} = \boldsymbol{L}^*\boldsymbol{L}+\boldsymbol{I}, \qquad \|\boldsymbol{K}\|_{\mathrm{op}}^2 = \kappa_K.$$

Assumption 3.2 implies that G is μ-weakly convex: $G(\boldsymbol{Y}) + \mu\|\boldsymbol{Y}\|_{\mathrm{F}}^2/2$ is convex. Since the theorem requires $\tau > \mu$, the joint $\boldsymbol{Y}$-update is uniquely defined.

Let $(\boldsymbol{\Sigma}^\star, \boldsymbol{Y}^\star, \boldsymbol{\Pi}^\star)$ be a split KKT point, whose existence is assumed in the theorem. Thus

$$\boldsymbol{\Sigma}^\star - \boldsymbol{S} + \boldsymbol{K}^*\boldsymbol{\Pi}^\star = 0, \qquad \boldsymbol{\Pi}^\star \in \partial G(\boldsymbol{Y}^\star), \qquad \boldsymbol{K}\boldsymbol{\Sigma}^\star = \boldsymbol{Y}^\star.$$

For $t \geq 1$, the exact ADMM updates yield

$$\begin{aligned} \mathbf{0} &= \boldsymbol{\Sigma}^{t+1} - \boldsymbol{S} + \boldsymbol{K}^*\boldsymbol{\Pi}^{t+1} + \tau\boldsymbol{K}^*(\boldsymbol{Y}^{t+1}-\boldsymbol{Y}^t), \\ \boldsymbol{\Pi}^{t+1} &\in \partial G(\boldsymbol{Y}^{t+1}), \\ \boldsymbol{\Pi}^{t+1} - \boldsymbol{\Pi}^t &= \tau(\boldsymbol{K}\boldsymbol{\Sigma}^{t+1} - \boldsymbol{Y}^{t+1}). \end{aligned}$$

The restriction $t \geq 1$ is needed because the second relation is assured only after the first joint $\boldsymbol{Y}$-update. Set

$$e_{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}^{t+1}-\boldsymbol{\Sigma}^\star, \quad e_{\boldsymbol{Y}} = \boldsymbol{Y}^{t+1}-\boldsymbol{Y}^\star, \quad e_{\boldsymbol{\Pi}} = \boldsymbol{\Pi}^{t+1}-\boldsymbol{\Pi}^\star, \quad d_{\boldsymbol{Y}} = \boldsymbol{Y}^{t+1}-\boldsymbol{Y}^t, \quad d_{\boldsymbol{\Pi}} = \boldsymbol{\Pi}^{t+1}-\boldsymbol{\Pi}^t,$$

and define

$$H^t = \frac{\tau}{2}\left\|\boldsymbol{Y}^t - \boldsymbol{Y}^\star\right\|_{\mathrm{F}}^2 + \frac{1}{2\tau}\left\|\boldsymbol{\Pi}^t - \boldsymbol{\Pi}^\star\right\|_{\mathrm{F}}^2.$$

Subtracting the KKT equations from the first ADMM optimality condition and using $\boldsymbol{K}e_{\boldsymbol{\Sigma}} = e_{\boldsymbol{Y}} + d_{\boldsymbol{\Pi}}/\tau$ gives the exact identity

$$H^t - H^{t+1} = \|e_{\boldsymbol{\Sigma}}\|_{\mathrm{F}}^2 + \langle e_{\boldsymbol{P}\boldsymbol{i}}, e_{\boldsymbol{Y}}\rangle + \langle d_{\boldsymbol{\Pi}}, d_{\boldsymbol{Y}}\rangle + \frac{\tau}{2}\|d_{\boldsymbol{Y}}\|_{\mathrm{F}}^2 + \frac{1}{2\tau}\|d_{\boldsymbol{\Pi}}\|_{\mathrm{F}}^2.$$

Weak monotonicity of ∂G gives

$$\langle e_{\boldsymbol{\Pi}}, e_{\boldsymbol{Y}}\rangle \geq -\mu\|e_{\boldsymbol{Y}}\|_{\mathrm{F}}^2, \qquad \langle d_{\boldsymbol{\Pi}}, d_{\boldsymbol{Y}}\rangle \geq -\mu\|d_{\boldsymbol{Y}}\|_{\mathrm{F}}^2.$$

Using $e_{\boldsymbol{Y}} = \boldsymbol{K}e_{\boldsymbol{\Sigma}} - d_{\boldsymbol{\Pi}}/\tau$ and Young's inequality with the η in the theorem therefore yields

$$\begin{aligned} H^t - H^{t+1} \geq{}& (1-\mu\kappa_K-\eta)\|e_{\boldsymbol{\Sigma}}\|_{\mathrm{F}}^2 + (\tau/2-\mu)\|d_{\boldsymbol{Y}}\|_{\mathrm{F}}^2 \\ &+ \frac{\tau/2-\mu-\mu^2\kappa_K/\eta}{\tau^2}\|d_{\boldsymbol{\Pi}}\|_{\mathrm{F}}^2. \end{aligned}$$

All coefficients are positive by assumption. Telescoping shows that $\boldsymbol{\Sigma}^t \to \boldsymbol{\Sigma}^\star$, $d_{\boldsymbol{Y}} \to 0$, and $d_{\boldsymbol{\Pi}} \to 0$, while the potential bounds $\boldsymbol{Y}^t$ and $\boldsymbol{\Pi}^t$. The dual update gives $\boldsymbol{K}\boldsymbol{\Sigma}^t - \boldsymbol{Y}^t \to 0$, so every accumulation point $(\boldsymbol{\Sigma}^\star, \overline{\boldsymbol{Y}}, \overline{\boldsymbol{\Pi}})$ is primal feasible. Passing to the limit in the ADMM optimality conditions, and using the closed graph of the limiting subdifferential of the weakly convex function G, shows that each such accumulation point is a split KKT point. Re-centering the same potential H^t at one accumulation point makes it nonincreasing and zero along the corresponding subsequence; hence it converges to zero along the entire sequence. Therefore $\boldsymbol{Y}^t \to \overline{\boldsymbol{Y}}$ and $\boldsymbol{\Pi}^t \to \overline{\boldsymbol{\Pi}}$. Expanding the limiting split KKT system shows

$$\mathbf{0} \in \boldsymbol{\Sigma}^\star - \boldsymbol{S} + \boldsymbol{L}^\top\partial P_\lambda(\boldsymbol{L}\boldsymbol{\Sigma}^\star) + \mathcal{N}_{\mathbb{S}_+(p)}(\boldsymbol{\Sigma}^\star),$$

which is exactly equation 7. Thus the ADMM limit is a stationary point of Problem 1.

C Proof of Statistical Properties

This section proves Theorem 5.3 and Corollary 5.4. We first control the sample-covariance noise on the oracle subspace and its orthogonal complement. We then combine these bounds with weak convexity and the PSD normal-cone condition.

Lemma C.1 (Projection noise). *Let $\mathcal{H}_0 \subseteq \mathbb{S}(p)$ be a fixed subspace of dimension r. Under Assumption 5.1,*

$$\|\Pi_{\mathcal{H}_0}\boldsymbol{\varepsilon}\|_{\mathrm{F}} \leq c_3\left(\sqrt{\frac{r}{n}} + \frac{t}{\sqrt{n}}\right)$$

with probability at least

$$1 - 2\exp\{-c_4 \min(t^2, t\sqrt{n})\}.$$

In particular,

$$\mathbb{E}\,\|\Pi_{\mathcal{H}_0}\boldsymbol{\varepsilon}\|_{\mathrm{F}}^2 \le c_3^2 \frac{r}{n}.$$

Proof. Choose an orthonormal basis of $\mathcal{H}_0$ and identify the projection with an r-dimensional random vector. For a unit $\boldsymbol{H} \in \mathcal{H}_0$,

$$\left\langle \boldsymbol{H}, \boldsymbol{x}_r \boldsymbol{x}_r^\top - \boldsymbol{\Sigma}^\star \right\rangle = \boldsymbol{x}_r^\top \boldsymbol{H} \boldsymbol{x}_r - \mathbb{E}[\boldsymbol{x}^\top \boldsymbol{H} \boldsymbol{x}]$$

is a centered sub-exponential random variable by the Hanson–Wright inequality. A vector Bernstein bound, or equivalently an ϵ-net argument on the unit sphere of $\mathcal{H}_0$, gives the stated two-regime tail for the average of these random vectors. Summing the second moments over the orthonormal basis gives the expectation bound. The important point is that this is a fixed-subspace bound; it does not union-bound over all matrix entries. □

Lemma C.2 (Green-scale dual noise)**.** *Define*

$$\boldsymbol{w} = \left(\boldsymbol{L}_{\mathcal{P}\setminus\mathcal{A}} \boldsymbol{L}_{\mathcal{P}\setminus\mathcal{A}}^\top\right)^+ \boldsymbol{L}_{\mathcal{P}\setminus\mathcal{A}} \boldsymbol{\varepsilon}.$$

Then

$$\boldsymbol{L}_{\mathcal{P}\setminus\mathcal{A}}^\top \boldsymbol{w} = \Pi_{\mathcal{U}_{\mathcal{A}}^\perp} \boldsymbol{\varepsilon}.$$

If $n \ge c_0 \log p$, then

$$\|\boldsymbol{w}\|_\infty \le C g_{\mathcal{A}} \sqrt{\frac{2\log p}{n}}$$

with probability at least $1 - c_2 p^{-2}$.

Proof. For $q \in \mathcal{P} \setminus \mathcal{A}$, write

$$\boldsymbol{M}_q = \boldsymbol{L}_{\mathcal{P}\setminus\mathcal{A}}^+ \boldsymbol{e}_q.$$

The Moore–Penrose identities imply

$$w_q = \langle \boldsymbol{\varepsilon}, \boldsymbol{M}_q \rangle .$$

By definition, $\|\boldsymbol{M}_q\|_{\mathrm{F}} \le g_{\mathcal{A}}$. Moreover,

$$w_q = \frac{1}{n} \sum_{r=1}^{n} \left[\boldsymbol{x}_r^\top \boldsymbol{M}_q \boldsymbol{x}_r - \mathbb{E}\{\boldsymbol{x}^\top \boldsymbol{M}_q \boldsymbol{x}\}\right].$$

Hanson–Wright followed by Bernstein's inequality gives

$$\mathbb{P}\{|w_q| > u\} \le 2\exp\left\{-cn \min\left(\frac{u^2}{g_{\mathcal{A}}^2}, \frac{u}{g_{\mathcal{A}}}\right)\right\}.$$

Set u proportional to $g_{\mathcal{A}}\sqrt{\log p / n}$ and use $n \gtrsim \log p$. A union bound over at most p^2 coordinates proves the result. □

Lemma C.3 (Penalty calculus)**.** *Under Assumption 3.2, for $s, t \in \mathbb{R}$ and $u \ge 0$,*

$$\rho_\lambda(|s|) - \rho_\lambda(|t|) \le \rho'_\lambda(|t|)(|s| - |t|) + \frac{\mu}{2}(|s| - |t|)^2,$$

$$\rho_\lambda(u) \ge \lambda u - \frac{\mu}{2} u^2, \qquad \rho_\lambda(u) \le \lambda u.$$

Consequently, for any $T \subseteq \mathcal{P}$,

$$2P_\lambda(\boldsymbol{L}\boldsymbol{\Delta}) + 2P_\lambda(\boldsymbol{L}\boldsymbol{\Sigma}) - 2P_\lambda(\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}) \le 4P_{\lambda,T^c}(\boldsymbol{L}\boldsymbol{\Sigma}) + 4P_{\lambda,T}(\boldsymbol{L}\boldsymbol{\Delta}),$$

where $\boldsymbol{\Delta} = \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$.

Proof. The first inequality follows from convexity of $\rho_\lambda(t) + \mu t^2/2$. The lower bound follows by supporting this convex function at zero, and the upper bound follows from concavity and $\rho_\lambda(0) = 0$. Subadditivity of a nonnegative concave function vanishing at zero gives the last display coordinate-wise. On T^c, use $\rho(|\widehat{z}|) \le \rho(|z|) + \rho(|\widehat{z} - z|)$; on T, use $\rho(|z|) \le \rho(|\widehat{z}|) + \rho(|\widehat{z} - z|)$. □

Lemma C.4 (Master inequality under the PSD constraint)**.** *Let $\widehat{\boldsymbol{\Sigma}}$ be a first order stationary point and $\boldsymbol{\Sigma} \in \mathbb{S}_+(p)$. Then*

$$\left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\|_{\mathrm{F}}^2 + \left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}^2 \leq \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star\|_{\mathrm{F}}^2 + 2\left\langle \boldsymbol{\varepsilon}, \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\rangle + 2P_\lambda(\boldsymbol{L}\boldsymbol{\Sigma}) - 2P_\lambda(\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}) + \mu\left\|\boldsymbol{L}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\right\|_2^2.$$

Proof. The first order stationary point condition gives a penalty subgradient $\boldsymbol{g} \in \partial(P_\lambda \circ \boldsymbol{L})(\widehat{\boldsymbol{\Sigma}})$ and a normal $\boldsymbol{n} \in \mathcal{N}_{\mathbb{S}_+(p)}(\widehat{\boldsymbol{\Sigma}})$ such that

$$\widehat{\boldsymbol{\Sigma}} - \boldsymbol{S} + \boldsymbol{g} + \boldsymbol{n} = \boldsymbol{0}.$$

Taking the inner product with $\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$, the normal-cone term is nonnegative after moving it to the appropriate side because $\mathbb{S}_+(p)$ is a closed convex cone. The weak-convex representation in Assumption 3.2 gives

$$-\left\langle \boldsymbol{g}, \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}\right\rangle \leq P_\lambda(\boldsymbol{L}\boldsymbol{\Sigma}) - P_\lambda(\boldsymbol{L}\widehat{\boldsymbol{\Sigma}}) + \frac{\mu}{2}\left\|\boldsymbol{L}(\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma})\right\|_2^2.$$

Expanding the quadratic loss by polarization and multiplying by two gives equation C.4. □

Proof of Theorem 5.3. Let $\boldsymbol{\Delta} = \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}$, and decompose

$$\boldsymbol{\varepsilon} = \Pi_{\mathcal{U}_\mathcal{A}}\boldsymbol{\varepsilon} + \Pi_{\mathcal{U}_\mathcal{A}^\perp}\boldsymbol{\varepsilon}.$$

On the event in Lemma C.2,

$$\begin{aligned} 2\langle \boldsymbol{\varepsilon}, \boldsymbol{\Delta}\rangle &\leq 2\left\|\Pi_{\mathcal{U}_\mathcal{A}}\boldsymbol{\varepsilon}\right\|_{\mathrm{F}}\|\boldsymbol{\Delta}\|_{\mathrm{F}} + 2\sum_{q\notin\mathcal{A}}|w_q||(\boldsymbol{L}\boldsymbol{\Delta})_q| \\ &\leq 2\left\|\Pi_{\mathcal{U}_\mathcal{A}}\boldsymbol{\varepsilon}\right\|_{\mathrm{F}}\|\boldsymbol{\Delta}\|_{\mathrm{F}} + \sum_{q\notin\mathcal{A}}\lambda|(\boldsymbol{L}\boldsymbol{\Delta})_q|. \end{aligned}$$

The lower bound in Lemma C.3 changes the last term into

$$P_{\lambda,\mathcal{A}^c}(\boldsymbol{L}\boldsymbol{\Delta}) + \frac{\mu}{2}\|\boldsymbol{L}\boldsymbol{\Delta}\|_2^2.$$

Combining this with the master inequality and equation C.3 yields

$$\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 + \left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}^2 \leq \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star\|_{\mathrm{F}}^2 + 4P_{\lambda,T^c}(\boldsymbol{L}\boldsymbol{\Sigma}) + 4P_{\lambda,T}(\boldsymbol{L}\boldsymbol{\Delta}) + 2a\|\boldsymbol{\Delta}\|_{\mathrm{F}} + \frac{3}{2}\mu\|\boldsymbol{L}\|_{\mathrm{op}}^2\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2,$$

where $a = \left\|\Pi_{\mathcal{U}_\mathcal{A}}\boldsymbol{\varepsilon}\right\|_{\mathrm{F}}$. For the selected coordinates,

$$\begin{aligned} P_{\lambda,T}(\boldsymbol{L}\boldsymbol{\Delta}) &\leq \lambda\|(\boldsymbol{L}\boldsymbol{\Delta})_T\|_1 \\ &\leq \lambda\sqrt{|T|}\,\|\boldsymbol{L}\|_{\mathrm{op}}\|\boldsymbol{\Delta}\|_{\mathrm{F}}. \end{aligned}$$

Consequently, after moving the weak-concavity term to the left and using $\|\boldsymbol{L}\|_{\mathrm{op}} \leq C_L$,

$$\begin{aligned} &\left(1 - \frac{3}{2}\mu C_L^2\right)\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 + \left\|\widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star\right\|_{\mathrm{F}}^2 \\ &\quad\leq \|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star\|_{\mathrm{F}}^2 + 4P_{\lambda,T^c}(\boldsymbol{L}\boldsymbol{\Sigma}) + \left(2a + 4C_L\lambda\sqrt{|T|}\right)\|\boldsymbol{\Delta}\|_{\mathrm{F}}. \end{aligned}$$

Set $\eta = 1 - \frac{3}{2}\mu C_L^2 > 0$ and $b = 4C_L\lambda\sqrt{|T|}$. Young's inequality gives

$$\begin{aligned} (2a+b)\|\boldsymbol{\Delta}\|_{\mathrm{F}} &\leq \frac{\eta}{2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 + \frac{(2a+b)^2}{2\eta} \\ &\leq \frac{\eta}{2}\|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 + \frac{4a^2+b^2}{\eta}. \end{aligned}$$

Dropping the remaining nonnegative $\eta \|\boldsymbol{\Delta}\|_{\mathrm{F}}^2 / 2$ therefore gives exactly the constants 4 and 16 in $\mathfrak{R}_\lambda$ and in the projected-noise term:

$$\left\| \widehat{\boldsymbol{\Sigma}} - \boldsymbol{\Sigma}^\star \right\|_{\mathrm{F}}^2 \leq \frac{\|\boldsymbol{\Sigma} - \boldsymbol{\Sigma}^\star\|_{\mathrm{F}}^2 + 4P_{\lambda,T^c}(\boldsymbol{L}\boldsymbol{\Sigma}) + 16C_L^2\lambda^2|T| + 4a^2}{\eta}.$$

Taking the infimum over $\boldsymbol{\Sigma} \in \mathbb{S}_+(p)$ and $T \subseteq \mathcal{P}$ gives equation 9. □

Proof of Corollary 5.4. Use $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}^\star$ and $T = \mathcal{A}$ in Theorem 5.3. Then $P_{\lambda,\mathcal{A}^c}(\boldsymbol{L}\boldsymbol{\Sigma}^\star) = 0$, and the selected-coordinate term is

$$C_L^2\lambda^2 K \lesssim \frac{C_L^2\zeta^2 K \log p}{n}.$$

Lemma C.1, applied to the fixed subspace $\mathcal{U}_\mathcal{A}$ of dimension $\overline{K}$, gives

$$\|\Pi_{\mathcal{U}_\mathcal{A}}\boldsymbol{\varepsilon}\|_{\mathrm{F}}^2 \lesssim \frac{\overline{K} + \log p}{n}.$$

Substitution proves equation 10. □

C.1 Nonconvex Penalty SCAD and MCP

For SCAD with shape parameter $\gamma > 2$,

$$\rho_\lambda(z) = \lambda \int_0^{|z|} \min\left\{1, \frac{(\gamma - t/\lambda)_+}{\gamma - 1}\right\} dt.$$

The corresponding μ in Assumption 3.2 is $\mu = \frac{1}{\gamma - 1}$. Its proximal operator is

$$\mathrm{prox}_{\rho_\lambda}(r) = \begin{cases} 0, & |r| \leq \lambda, \\ \mathrm{sign}(r)(|r| - \lambda), & \lambda < |r| \leq 2\lambda, \\ \mathrm{sign}(r)\dfrac{(\gamma - 1)|r| - \gamma\lambda}{\gamma - 2}, & 2\lambda < |r| \leq \gamma\lambda, \\ r, & |r| > \gamma\lambda. \end{cases}$$

For MCP with shape parameter $\gamma > 1$,

$$\rho_\lambda(z) = \lambda \int_0^{|z|} \left(1 - \frac{t}{\gamma\lambda}\right)_+ dt.$$

The corresponding μ in Assumption 3.2 is $\mu = \frac{1}{\gamma}$. Its proximal operator is

$$\mathrm{prox}_{\rho_\lambda}(r) = \begin{cases} 0, & |r| \leq \lambda, \\ \mathrm{sign}(r)\dfrac{\gamma(|r| - \lambda)}{\gamma - 1}, & \lambda < |r| \leq \gamma\lambda, \\ r, & |r| > \gamma\lambda. \end{cases}$$