Supplementary Material
Principal Component Hierarchy for Sparse Quadratic Programs

A Additional Numerical Experiments

A.1 Comparison of Computational Time to `warm start`

We study the impact of the sample size N on the recovery quality of the solution. We fix $n = 1000$, $s = 10$, $\rho = 0.5$, $\text{SNR} = 6$ and $\eta = 10$. We showcase the computational time of our methods and of the `warm start` in Figure A, the computational time is defined as the time needed to generate $x^\star$. Note that the `BR` method uses MOSEK to obtain the solution to $x^\star$ because it does not converge to a single set z for $\eta = 10$, so the solver time is also included in the computational time. We run the `BR` method for $T_{\texttt{BR}} = 20$ iterations, and we run the `DP` method for $T_{\texttt{DP}} = 500$ iterations.

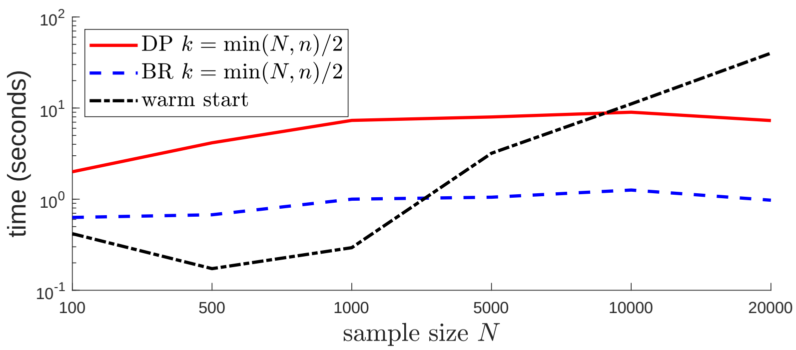

Figure A: Computational time over different sample sizes averaged over 25 replications

We observe that the computational time of the `DP` method increases monotonically with the sample size N. Note that $T_{\texttt{BR}} \ll T_{\texttt{DP}}$ so calculating $Z_{\texttt{BR}}$ requires less time than $Z_{\texttt{DP}}$. We observe that when $N = 100$ the `BR` and the `warm start` have a higher computational time than for $N = 500$. For the `BR`, this is because the number of non-zero elements in Z (i.e., $\|Z\|_0$) is larger for $N = 100$ than for $N = 500$, hence MOSEK takes more time for $N = 100$. The MSE of all methods is similar when $N \geq 500$, when $N = 100$ the MSE of all methods differs significantly at every instance. This is also observed by Bertsimas & van Parys (2017), which states that the computational time and MSE deteriorate as N gets smaller relative to n.
We observe that `BR` and `DP` perform particularly well in terms of computational time in ranges where $N > n$ compared to the `warm start`. The running time of our method is less susceptive to the number of samples N. This is in stark contrast to the `warm start`, in which the kernel matrix of dimension N-by-N is stored.

A.2 Comparison for Different SNR and s

We extend the comparison made in the paper for different values of SNR and s. In Table A and B we observe that the MSE over different SNR and s is very similar for all methods. This is due to the fact that all methods find a

Table A: MSE over different SNR averaged over 25 independent replications. Lower is better.

	`DP` $k = 400$	`BR` $k = 400$	`warm start`	`Beck Alg 7`	`KDD`
SNR = 20	0.588	0.588	0.588	0.588	0.588
SNR = 6	1.767	1.767	1.767	1.767	1.767
SNR = 3	3.452	3.452	3.452	3.452	3.452
SNR = 1	10.190	10.190	10.190	10.198	10.205
SNR = 0.05	194.592	194.561	194.560	194.756	199.928

Table B: MSE over different s averaged over 25 independent replications. Lower is better.

	`DP` $k = 400$	`BR` $k = 400$	`warm start`	`Beck Alg 7`	`KDD`
$s = 5$	0.887	0.887	0.887	0.887	0.887
$s = 10$	1.767	1.767	1.767	1.767	1.767
$s = 20$	3.435	3.435	3.435	3.557	3.450
$s = 30$	5.050	5.050	5.058	5.888	5.440
$s = 40$	6.919	6.928	6.918	8.290	8.560

similar support $z^\star$. Using this support all problems solve the same convex quadratic programming problem. We also observe that the reduced size $\|Z_{\mathtt{BR}}\|_0 \approx 2s$ and $\|Z_{\mathtt{DP}}\|_0 \approx s$. So as the problem in $(\mathcal{P}_Z)$ increases with s, MOSEK takes more time to solve $(\mathcal{P}_Z)$ and because $\|Z_{\mathtt{BR}}\|_0 > \|Z_{\mathtt{DP}}\|_0$ the `DP` is faster for large s.

A.3 Real Datasets

For the real datasets listed in the main paper, we present the out-sample MSE for the different methods in Table C.

Table C: Out-sample MSE on real datasets, averaged over 50 independent train-test splits. Lowest error for each dataset is highlighted in grey.

	`DP` $k = 40$	`DP` $k = \hat{k}$	`BR` $k = 40$	`BR` $k = \hat{k}$	`warm start`	`screening`	`BH Alg 7`	`KDD`
(FB)	3.026×10^{-4}	3.025×10^{-4}	3.022×10^{-4}	$\mathbf{3.020 \times 10^{-4}}$	out of memory	3.022×10^{-4}	3.203×10^{-4}	3.409×10^{-4}
(ON)	$\mathbf{1.796 \times 10^{-4}}$	1.797×10^{-4}	1.797×10^{-4}	1.797×10^{-4}	1.797×10^{-4}	1.797×10^{-4}	1.803×10^{-4}	1.803×10^{-4}
(SC)	1.263×10^{-2}	1.263×10^{-2}	1.398×10^{-2}	1.370×10^{-2}	1.326×10^{-2}	$\mathbf{1.257 \times 10^{-2}}$	1.454×10^{-2}	1.473×10^{-2}
(CR)	2.892×10^{-2}	2.891×10^{-2}	2.893×10^{-2}	2.894×10^{-2}	2.900×10^{-2}	$\mathbf{2.868 \times 10^{-2}}$	3.103×10^{-2}	3.148×10^{-2}
(UJ)	$\mathbf{2.149 \times 10^{-2}}$	2.324×10^{-2}	2.684×10^{-2}	2.691×10^{-2}	2.468×10^{-2}	2.291×10^{-2}	3.848×10^{-2}	3.080×10^{-2}

Similar to the in-sample MSE, Table C shows that `DP` delivers a lower out-sample MSE than `BR` in 4 out of 5 datasets, and `DP` also has a lower out-sample MSE than the `warm start`, `BH Alg 7` and `KDD` for all datasets. The `screening` method outperforms the `DP` on the (SC) and (CR) dataset, however as explained in the main paper for $\eta = \sqrt{N_{train}}$ the result of `screening` in Table C on the (SC) and (CR) datasets is essentially the results obtained by applying the MOSEK solver to the original problem (reaching a time limit of 300 seconds).

B Proof of Proposition 3.1

We provide the proof of Proposition 3.1, which is not included in the main paper.

Proof. Using the big-M equivalent formulation, we have

$$
\begin{aligned}
\mathcal{J}_k^\star = \min_{\substack{z \in \{0,1\}^n \\ \sum z_j \le s}} \quad \min \quad & \sum_{i=1}^{k} \lambda_i y_i^2 + \langle c, x\rangle + \eta^{-1}\|x\|_2^2 \\
\text{s.t.} \quad & x \in \mathbb{R}^n, \ y \in \mathbb{R}^k \\
& \sqrt{\lambda_i} y_i = \sqrt{\lambda_i}\, \langle v_i, x\rangle \qquad i \in [k] \\
& |x_j| \le M z_j \qquad j \in [n] \\
& Ax \le b.
\end{aligned}
$$

Fix a feasible solution for z and consider the inner minimization problem. By associating the first two constraints with the dual variables α and β, the Lagrangian function is defined as

$$\begin{aligned}\mathcal{L}(x, y, \alpha, \beta) &= \sum_{i=1}^{k} \lambda_i y_i^2 + \langle c, x\rangle + \eta^{-1}\|x\|_2^2 + \sum_{i=1}^{k} \alpha_i \sqrt{\lambda_i}\left(\langle v_i, x\rangle - y_i\right) + \beta^\top (Ax - b) \\ &= -\beta^\top b + y^\top \Lambda y - \alpha^\top \sqrt{\Lambda} y + \left\langle c + V\sqrt{\Lambda}\alpha + A^\top \beta, x\right\rangle + \eta^{-1}\|x\|_2^2,\end{aligned}$$

in which $\Lambda = \mathrm{diag}\{\lambda_1, \cdots, \lambda_k\}$. For any feasible solution z, the inner minimization problem is a convex quadratic optimization problem and we have

$$\mathcal{J}_k^\star = \min_{\substack{z\in\{0,1\}^n \\ \sum z_j \le s}} \max_{\substack{\alpha\in\mathbb{R}^k \\ \beta\in\mathbb{R}_+^m}} L(z, \alpha, \beta),$$

where the objective function L is defined as

$$L(z, \alpha, \beta) := -\beta^\top b + \min_{y\in\mathbb{R}^k} y^\top \Lambda y - \alpha^\top \sqrt{\Lambda} y + \min_{\substack{x\in\mathbb{R}^n \\ |x_j|\le M z_j \ \forall j}} \left\langle c + V\sqrt{\Lambda}\alpha + A^\top\beta, x\right\rangle + \eta^{-1}\|x\|_2^2.$$

We will reformulate the two optimization subproblems in the definition of L. For any feasible pair $\beta \in \mathbb{R}_+^m$ and $\alpha \in \mathbb{R}^k$, the subproblem over y is an unconstrained convex quadratic optimization problem. The corresponding optimal solution for y is

$$y^\star(\alpha, \beta) = \frac{1}{2}(\sqrt{\Lambda})^{-1}\alpha.$$

Consequently, the optimal value of the y-subproblem is given by

$$\min_{y\in\mathbb{R}^k} y^\top \Lambda y - \alpha^\top \sqrt{\Lambda} y = -\frac{1}{4}\|\alpha\|_2^2.$$

Next, consider the x-subproblem. Let $\gamma := c + V\sqrt{\Lambda}\alpha + A^\top\beta$ and let γ_j denote the j-th element of γ. The big-M equivalent formulation for the x-subproblem admits the form

$$\min_{\substack{x\in\mathbb{R}^n \\ |x_j|\le M z_j \ \forall j}} \sum_{j=1}^{n} \gamma_j x_j + \frac{x_j^2}{\eta} = \sum_{j=1}^{n} \min_{\substack{x\in\mathbb{R}^n \\ |x_j|\le M z_j \ \forall j}} \gamma_j x_j + \frac{x_j^2}{\eta} = \sum_{j=1}^{n} -\frac{\eta}{4}\gamma_j^2 z_j,$$

where the last equality exploits the fact that the optimal solution of x_j is

$$x_j^\star(z_j) = \begin{cases} -\frac{\eta}{2}\gamma_j & \text{if } z_j = 1, \\ 0 & \text{if } z_j = 0. \end{cases}$$

We thus have

$$L(z, \alpha, \beta) = -\beta^\top b - \frac{1}{4}\sum_{i=1}^{k} \alpha_i^2 - \sum_{j=1}^{n} \frac{\eta}{4}\gamma_j^2 z_j,$$

where $\gamma = c + V\sqrt{\Lambda}\alpha + A^\top\beta$ and γ_j is the j-th element of γ. Rewriting the summations using norm and matrix multiplications completes the proof. □

C Principal Component Hierarchy for Sparsity-Penalized Quadratic Programs

The approach proposed in the main paper can be extended to solve the $\|\cdot\|_0$-penalized problem of the form

$$\begin{array}{ll} \min & \langle c, x\rangle + \langle x, Qx\rangle + \eta^{-1}\|x\|_2^2 + \theta\|x\|_0 \\ \text{s.t.} & x \in \mathbb{R}^n,\ Ax \le b \end{array}$$

for some sparsity-inducing parameter $\theta > 0$. The corresponding approximation using k principal components of the matrix Q is

$$
\begin{aligned}
\mathcal{U}_k^\star \triangleq \quad \min \quad & \langle c, x\rangle + \sum_{i=1}^k \lambda_i y_i^2 + \eta^{-1}\|x\|_2^2 + \theta\|x\|_0 \\
\text{s.t.} \quad & x \in \mathbb{R}^n,\ y \in \mathbb{R}^k \\
& Ax \le b \\
& \sqrt{\lambda_i} y_i = \sqrt{\lambda_i}\,\langle v_i, x\rangle \quad i \in [k].
\end{aligned}
\tag{$\mathcal{W}_k$}
$$

Proposition C.1 (Min-max characterization). *For each $k \le n$, the optimal value of problem $(\mathcal{W}_k)$ is equal to*

$$
\mathcal{U}_k^\star = \min_{z\in\{0,1\}^n} \max_{\substack{\alpha\in\mathbb{R}^k \\ \beta\in\mathbb{R}_+^m}} H(z,\alpha,\beta),
$$

where the objective function H is defined as

$$
H(z,\alpha,\beta) \triangleq \theta\sum_{j=1}^n z_j - \beta^\top b - \frac{1}{4}\|\alpha\|_2^2 - \frac{\eta}{4}(c + V\sqrt{\Lambda}\alpha + A^\top\beta)^\top \mathrm{diag}(z)(c + V\sqrt{\Lambda}\alpha + A^\top\beta). \tag{C.1}
$$

Proof of Proposition C.1. The sparsity-penalized principal component approximation problem can be rewritten using the big-M formulation as

$$
\begin{aligned}
\min_{z\in\{0,1\}^n} \quad \min \quad & \langle c, x\rangle + \sum_{i=1}^k \lambda_i y_i^2 + \eta^{-1}\|x\|_2^2 + \theta\sum_{j=1}^n z_j \\
\text{s.t.} \quad & x \in \mathbb{R}^n,\ y \in \mathbb{R}^k \\
& \sqrt{\lambda_i} y_i = \sqrt{\lambda_i}\,\langle v_i, x\rangle \quad i \in [k] \\
& |x_j| \le M z_j \quad j \in [n] \\
& Ax \le b.
\end{aligned}
$$

For any feasible solution z, the inner minimization problem is a convex quadratic optimization problem. By strong duality, we have the equivalent problem

$$
\mathcal{U}_k^\star = \min_{z\in\{0,1\}^n} \max_{\substack{\alpha\in\mathbb{R}^k \\ \beta\in\mathbb{R}_+^m}} H(z,\alpha,\beta),
$$

where the objective function H is

$$
H(z,\alpha,\beta) = -\beta^\top b + \min_{y\in\mathbb{R}^k}\ y^\top\sqrt{\Lambda}y - \alpha^\top \mathrm{diag}(\sqrt{\Lambda})y + \min_{\substack{x\in\mathbb{R}^n \\ |x_j|\le Mz_j\ \forall j}} \left\langle c + V\mathrm{diag}(\sqrt{\Lambda})\alpha + A^\top\beta, x\right\rangle + \eta^{-1}\|x\|_2^2 + \theta\sum_{j=1}^n z_j.
$$

Following proposition 3.1 we can calculate the optimal values for $y^\star$ and $x^\star$. Considering the x-subproblem, let $\gamma = c + V\sqrt{\Lambda}\alpha + A^\top\beta$ and γ_j be the j-th element of γ. The big-M equivalent formulation for the x-subproblem admits the form

$$
\begin{aligned}
\min_{\substack{x\in\mathbb{R}^n \\ |x_j|\le Mz_j\ \forall j}} \sum_{j=1}^n \gamma_j x_j + \frac{x_j^2}{\eta} + \theta z_j &= \sum_{j=1}^n \min_{\substack{x\in\mathbb{R}^n \\ |x_j|\le Mz_j\ \forall j}} \gamma_j x_j + \frac{x_j^2}{\eta} + \theta z_j \\
&= \sum_{j=1}^n (-\frac{\eta}{4}\gamma_j^2 + \theta) z_j,
\end{aligned}
\tag{C.2}
$$

where the last equality exploits the fact that the optimal solution of x_j is

$$
x_j^\star(z_j) = \begin{cases} -\frac{\eta}{2}\gamma_j & \text{if } \frac{\eta}{4}\gamma_j^2 > \theta, \\ 0 & \text{if } \frac{\eta}{4}\gamma_j^2 \le \theta. \end{cases}
$$

As a consequence, we have

$$H(z,\alpha,\beta) = -\beta^\top b - \frac{1}{4}\sum_{i=1}^{k}\alpha_i^2 + \sum_{j=1}^{n}(-\frac{\eta}{4}\gamma_j^2 + \theta)z_j,$$

where $\gamma = c + V\sqrt{\Lambda}\alpha + A^\top\beta$ and γ_j is the j-th element of γ. Rewriting the summations using norm and matrix multiplications completes the proof. $\square$

Lemma C.2 (Closed-form minimizer)**.** *Given any pair* (α, β)*, the minimizer of the function* H *defined in* (C.1) *can be computed as*

$$\arg\min_{z\in\{0,1\}^n} H(z,\alpha,\beta) = \mathbb{I}\big\{\frac{\eta}{4}\mathrm{diag}((c+V\sqrt{\Lambda}\alpha+A^\top\beta)(c+V\sqrt{\Lambda}\alpha+A^\top\beta)^\top) > \theta\big\},$$

where $\mathbb{I}$ *is the component-wise indicator function and the* diag *operator here returns the vector of diagonal elements of the input matrix.*

This lemma immediately follows from (C.1).

References

Bertsimas, D. and van Parys, B. Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. *The Annals of Statistics*, 48:300–323, 2017.