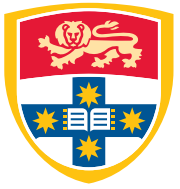
Robust estimation of precision matrices under cellwise contamination

Garth Tarr, Samuel Müller and Neville Weber

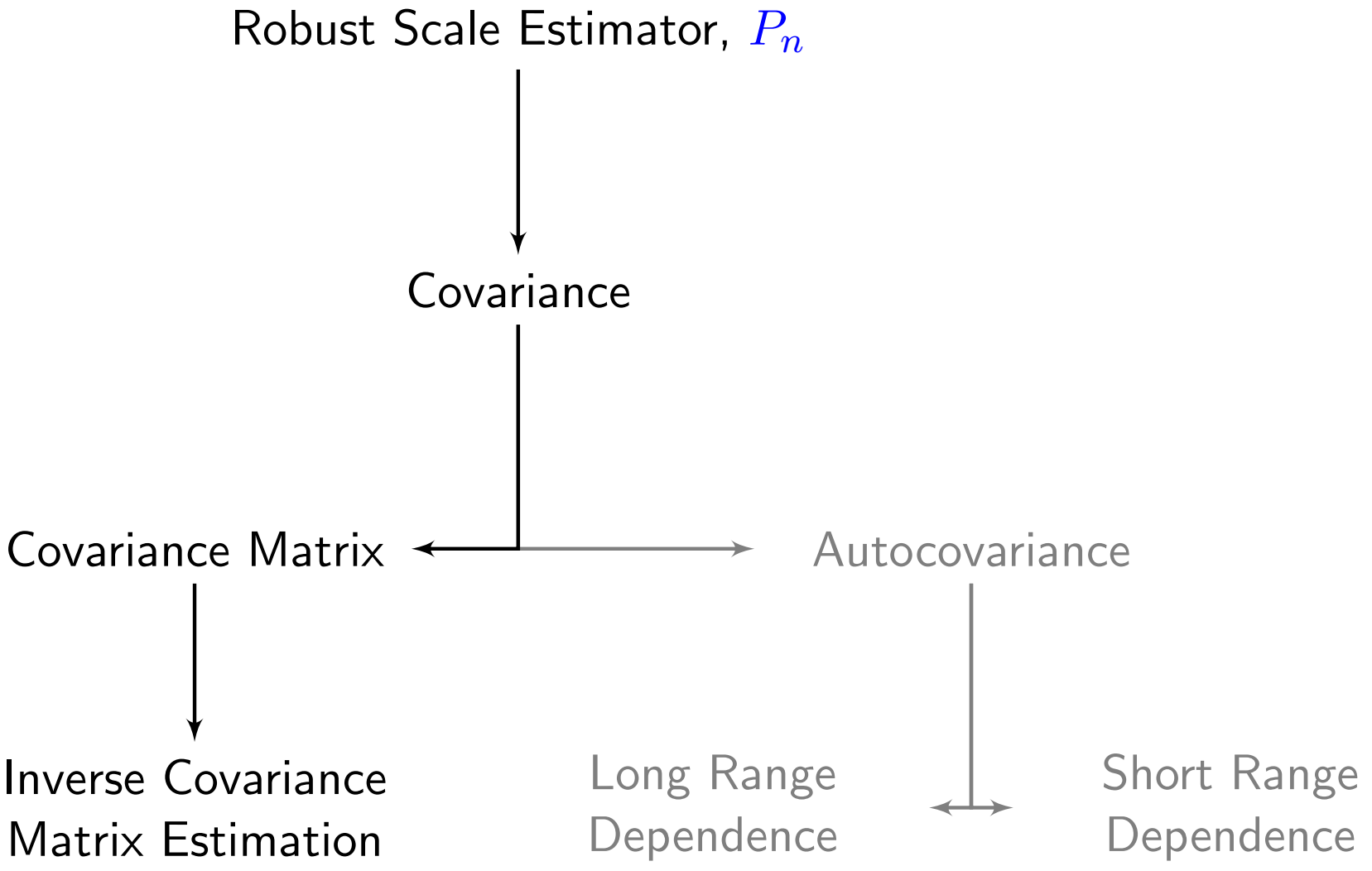


THE UNIVERSITY OF SYDNEY

Australian National University

COMPSTAT 2014

Outline


Robust Scale Estimator, P_n
Covariance
Covariance Matrix
Autocovariance
Inverse Covariance Matrix Estimation
Long Range Dependence
Short Range Dependence

Outline

Pairwise mean scale estimator: P_n

- Consider the U-statistic, based on the pairwise mean kernel,

$$U_n(\mathbf{X}) := \binom{n}{2}^{-1} \sum_{i<j} \frac{X_i + X_j}{2}.$$

- Let $H(t) = P((X_i + X_j)/2 \leq t)$ be the cdf of the kernels with corresponding empirical distribution function,

$$H_n(t) := \binom{n}{2}^{-1} \sum_{i<j} \mathbb{I}\left\{\frac{X_i + X_j}{2} \leq t\right\}, \quad \text{for } t \in \mathbb{R}.$$

Definition (Interquartile range of pairwise means)

$$P_n = c\left[H_n^{-1}(0.75) - H_n^{-1}(0.25)\right],$$

where $c \approx 1.048$ is a correction factor to ensure P_n is consistent for the standard deviation when the underlying observations are Gaussian.

Outline

From scale to covariance: the GK device

- Gnanadesikan and Kettenring (1972) relate scale and covariance using the following identity,

$$\mathrm{cov}(X, Y) = \frac{1}{4\alpha\beta}\left[\mathrm{var}(\alpha X + \beta Y) - \mathrm{var}(\alpha X - \beta Y)\right],$$

where X and Y are random variables.

- In general, X and Y can have different units, so we set $\alpha = 1/\sqrt{\mathrm{var}(X)}$ and $\beta = 1/\sqrt{\mathrm{var}(Y)}$.

- Replacing **variance** with $\boldsymbol{P_n^2}$ we can similarly construct,

$$\gamma_P(X, Y) = \frac{1}{4\alpha\beta}\left[P_n^2(\alpha X + \beta Y) - P_n^2(\alpha X - \beta Y)\right],$$

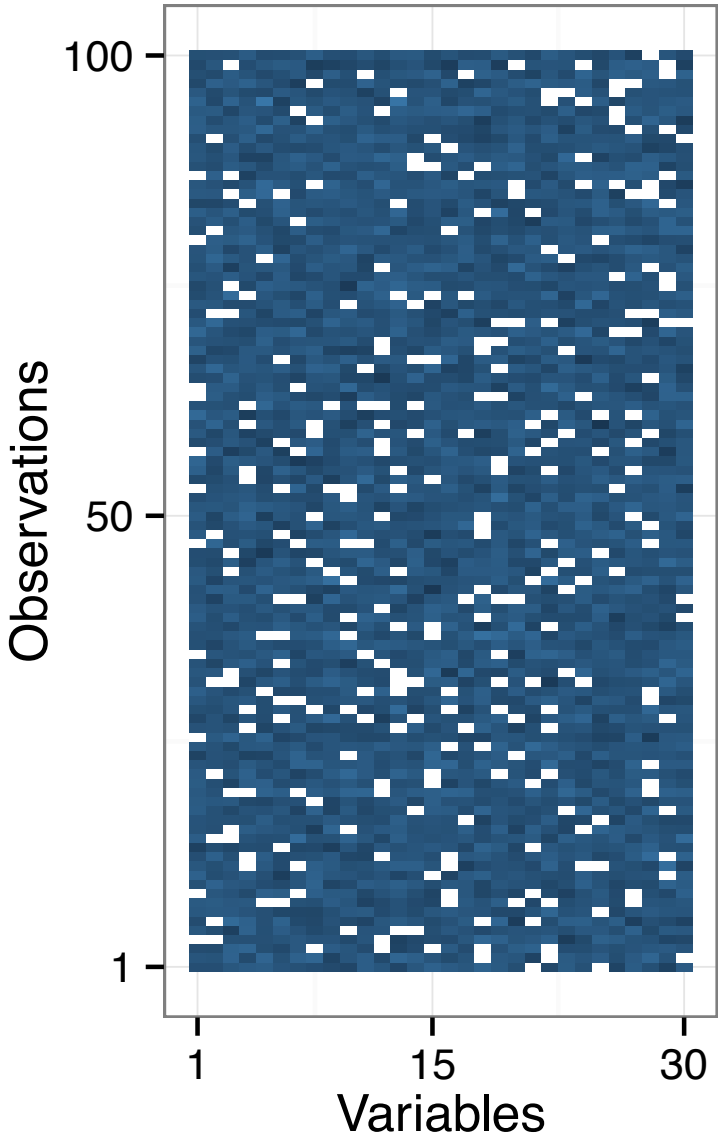
where $\alpha = 1/P_n(X)$ and $\beta = 1/P_n(Y)$.

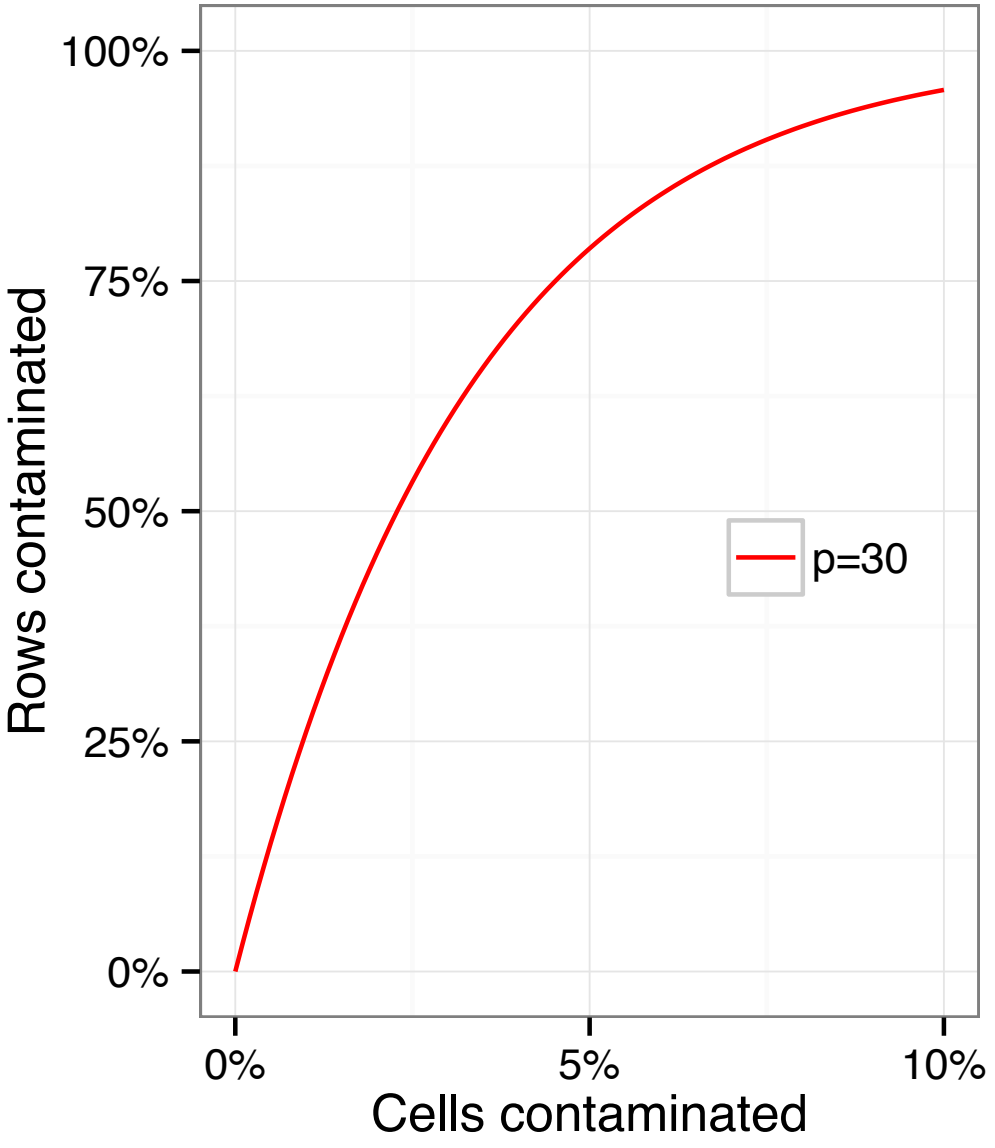
Outline

Estimating dependence

Problem:

To estimate dependence in multivariate settings with cellwise contamination.





For details see Alqallaf et al. (2009).

Estimating dependence

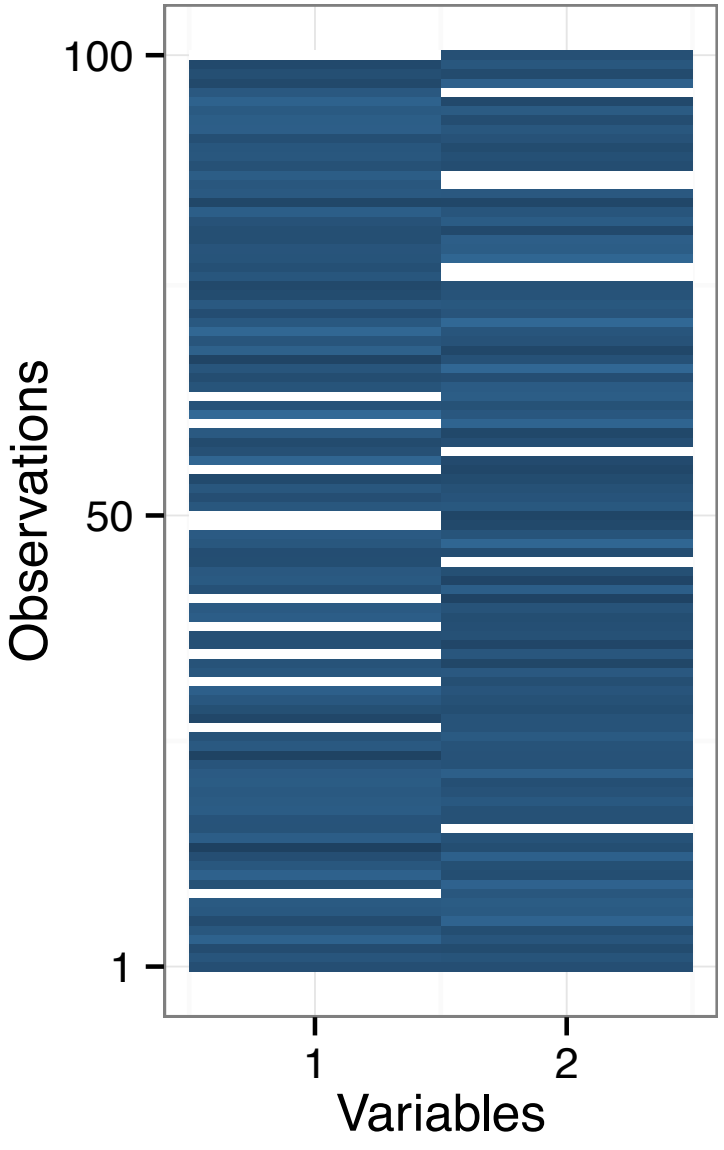
Problem:

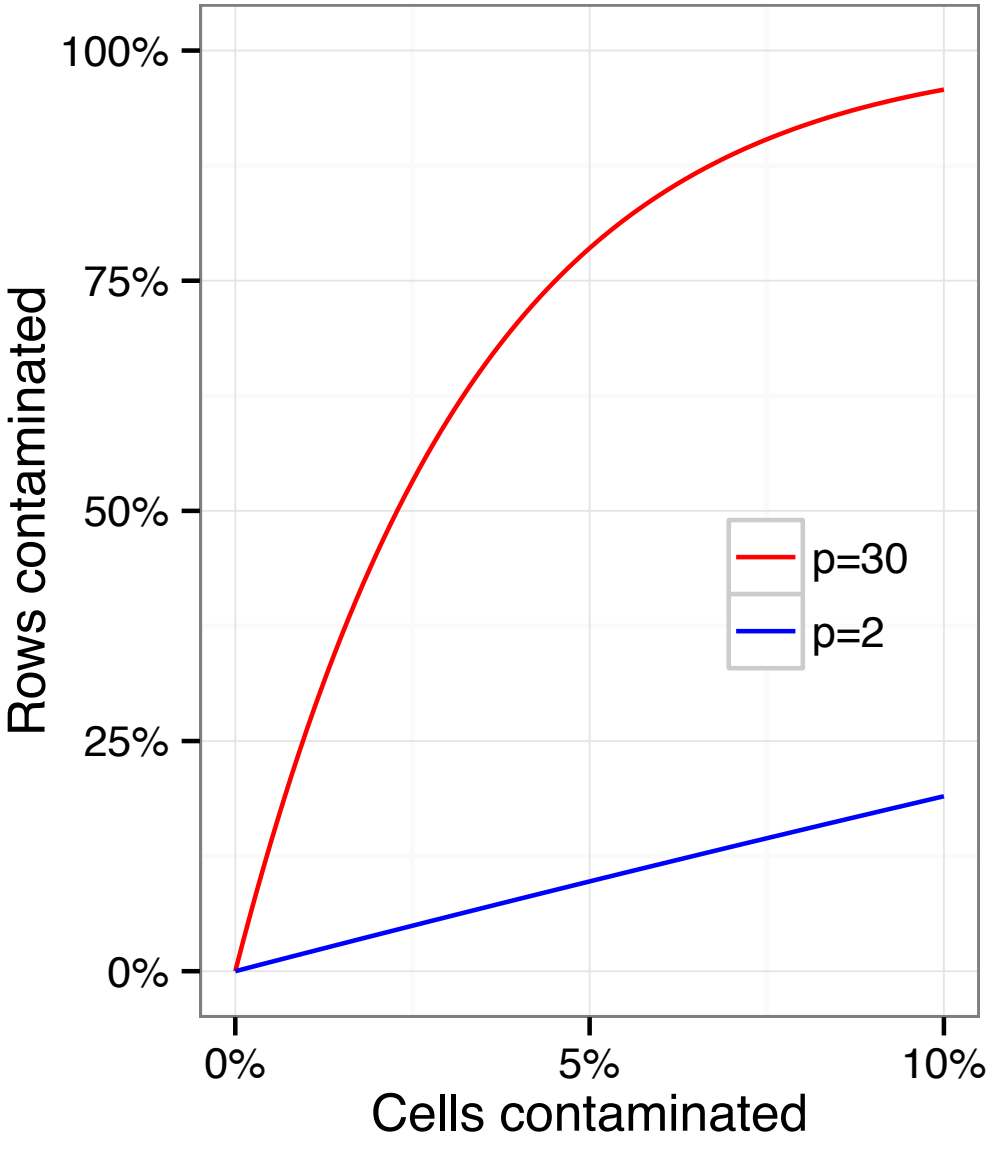
To estimate dependence in multivariate settings with cellwise contamination.

Solution:

1. pairwise covariance matrices
2.
3.

Pairwise approach to the rescue?





Estimating dependence

Problem:

To estimate dependence in multivariate settings with cellwise contamination.

Solution:

1. pairwise covariance matrices
2. correct for positive definiteness
3.

Positive definite?

- Standard approach of Maronna and Zamar (2002) suffers from outlier propagation so fails for cellwise contamination.

- Higham (2002) outlines the nearest positive definite (NPD) approach:
  1. Perform a spectral decomposition of the symmetric matrix of pairwise covariances
  2. Any negative eigenvalues are set to some small positive constant
  3. Reconstruct the covariance matrix using the adjusted eigenstructure

! NPD approach produces poorly conditioned covariance matrices.

Estimating dependence

Problem:

To estimate dependence in multivariate settings with cellwise contamination.

Solution:

1. pairwise covariance matrices
2. correct for positive definiteness
3. regularisation procedure

Precision matrices

- In many practical applications, the covariance matrix is not what is really required.
- PCA, Mahalanobis distance, LDA, etc. use the inverse covariance matrix: the precision matrix, $\mathbf{\Theta} = \mathbf{\Sigma}^{-1}$.
- Precision matrices are also of interest in modelling Gaussian Markov random fields, where zeros in the correspond to conditional independence between variables.

Sparsity!

In many applications, it is often useful to impose a level of sparsity on the estimated precision matrix.

Regularisation techniques

Graphical lasso (glasso) (Friedman, Hastie, Tibshirani, 2007) minimises the penalised negative Gaussian log-likelihood:

$$f(\mathbf{\Theta}) = \mathrm{tr}(\hat{\mathbf{\Sigma}}\mathbf{\Theta}) - \log|\mathbf{\Theta}| + \lambda||\mathbf{\Theta}||_1,$$

where $||\mathbf{\Theta}||_1$ is the L_1 norm and λ is a tuning parameter for the amount of shrinkage.

Quadratic Inverse Covariance (QUIC) (Hsieh, et. al. 2011) solves the same minimisation problem as the glasso but uses a second order approach.

Constrained ℓ_1-minimisation for inverse matrix estimation (CLIME) (Cai, Liu and Luo, 2011) solves the following objective function:

$$\min||\mathbf{\Theta}||_1 \quad \text{subject to: } ||\hat{\mathbf{\Sigma}}\mathbf{\Theta} - \mathbf{I}||_\infty \leq \lambda.$$

Estimating dependence

Problem:

To estimate dependence in multivariate settings with cellwise contamination.

Solution:

1. pairwise covariance matrices
2. correct for positive definiteness
3. regularisation procedure

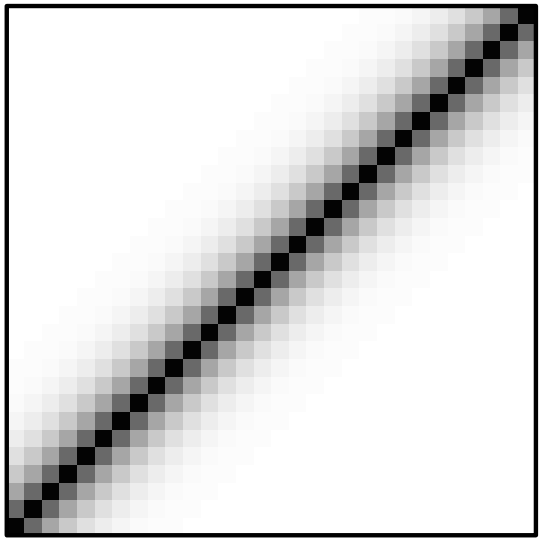
Evaluation:

- Is the estimate "close" to the truth?
-

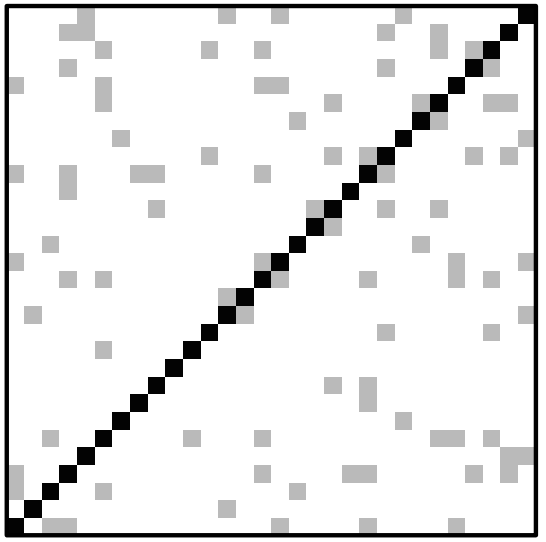
Simulation design

sample size	$n = 100$
replications	$N = 100$
variables	$p = 30, 60, 90$
scenarios	banded, sparse and dense precision matrices, $\mathbf{\Theta}$
true data	$\mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma} = \mathbf{\Theta}^{-1}$
contamination	0% to 25% randomly scattered component-wise

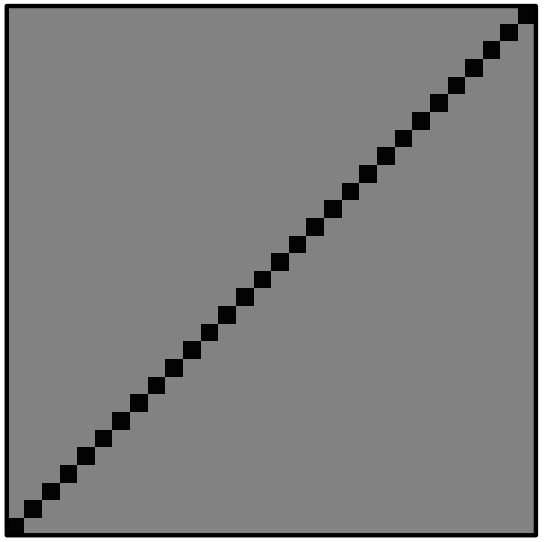
1. Banded



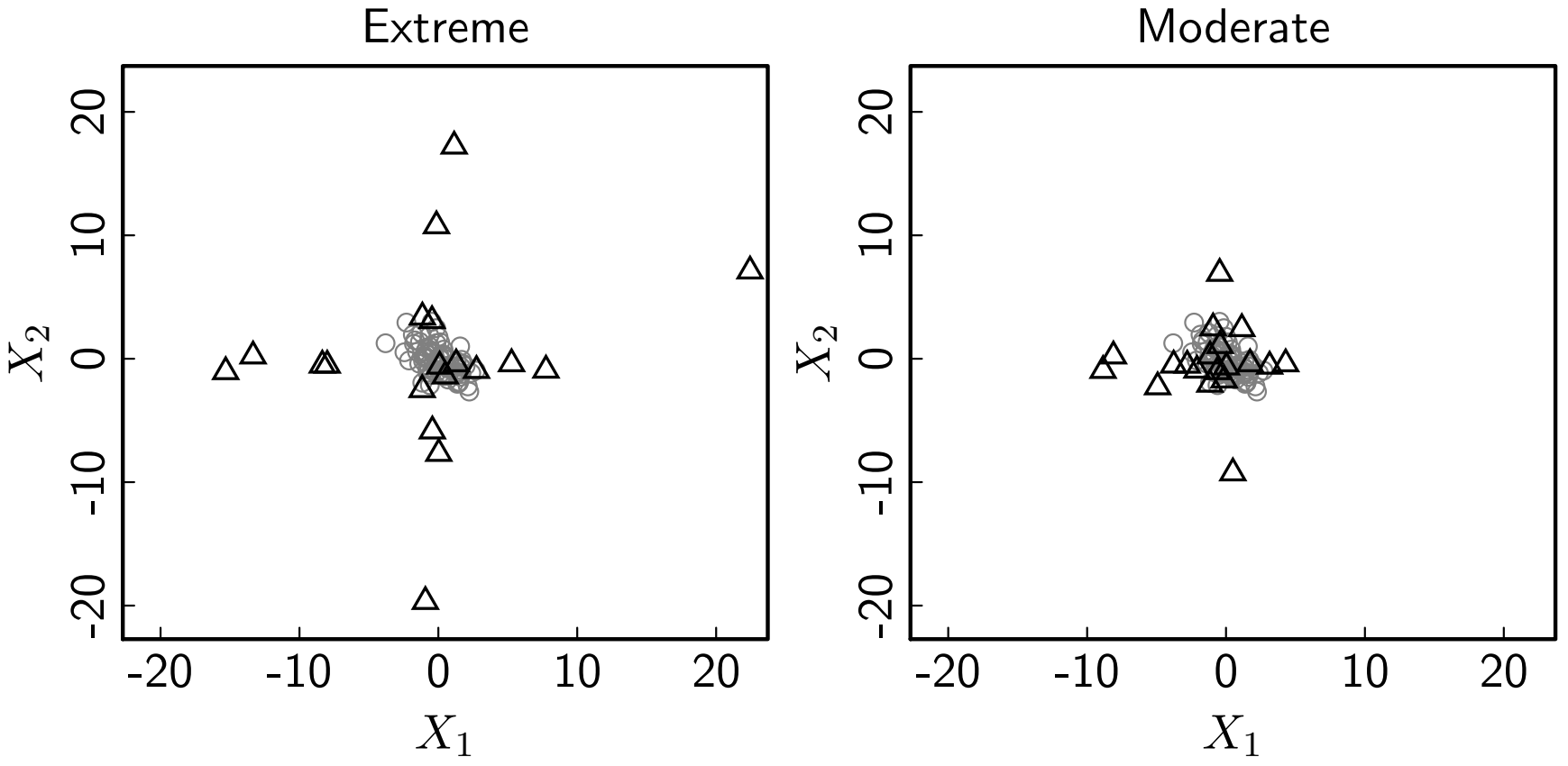
2. Sparse



3. Dense



10% contamination



Evaluating performance

Entropy Loss

Measures how "close" $\hat{\mathbf{\Theta}}$ is to $\mathbf{\Theta}$,

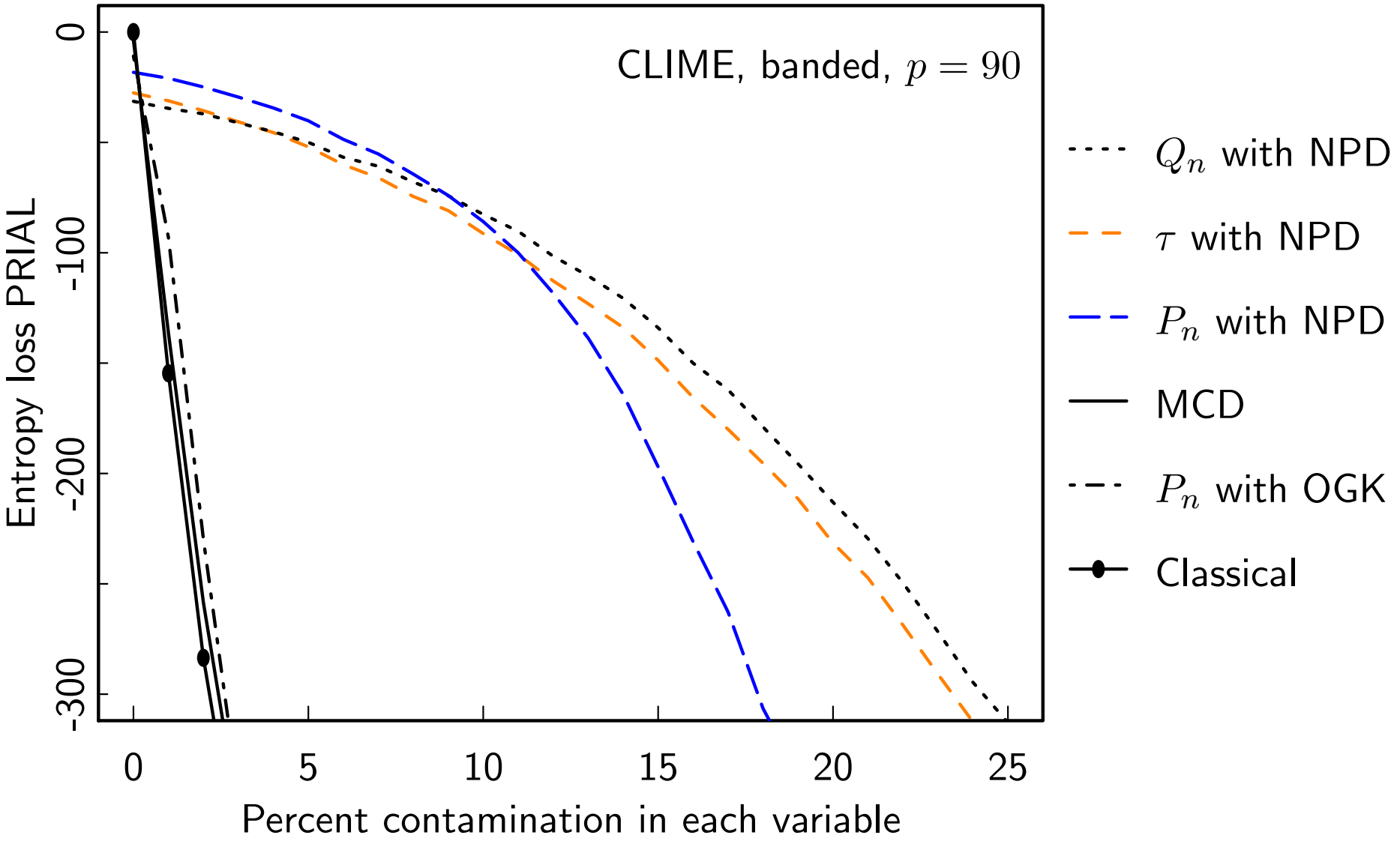
$$L(\mathbf{\Theta}, \hat{\mathbf{\Theta}}) = \operatorname{tr}(\mathbf{\Theta}^{-1}\hat{\mathbf{\Theta}}) - \log|\mathbf{\Theta}^{-1}\hat{\mathbf{\Theta}}| - p.$$

Reported as percentage relative improvement in average loss:

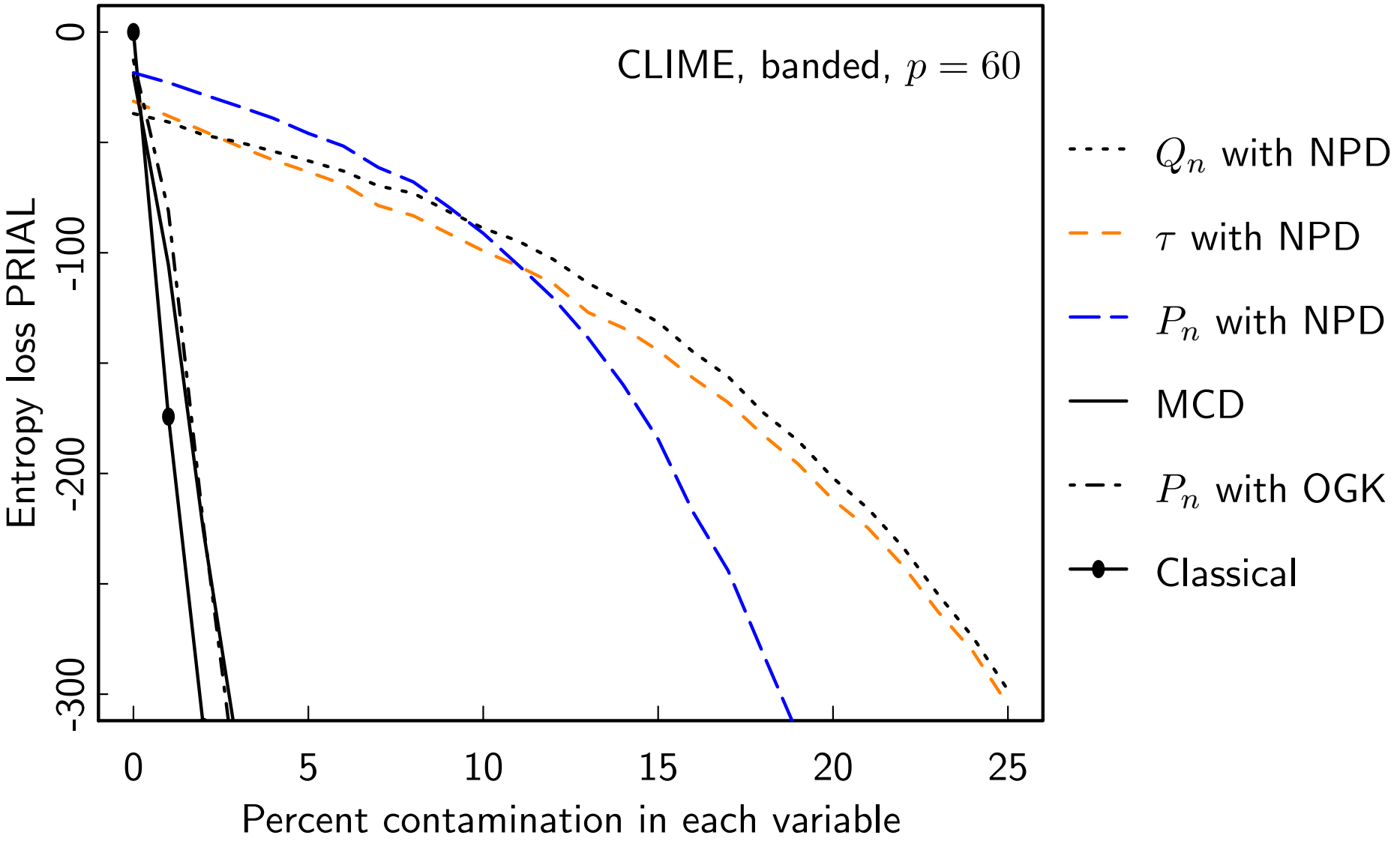
$$\text{PRIAL}(\hat{\mathbf{\Theta}}) = \frac{L(\mathbf{\Theta}, \hat{\mathbf{\Theta}}_0) - L(\mathbf{\Theta}, \hat{\mathbf{\Theta}})}{L(\mathbf{\Theta}, \hat{\mathbf{\Theta}}_0)} \times 100,$$

where $\hat{\mathbf{\Theta}}_0$ is the estimated precision matrix after a regularisation technique has been applied to the classical sample covariance matrix for uncontaminated data.

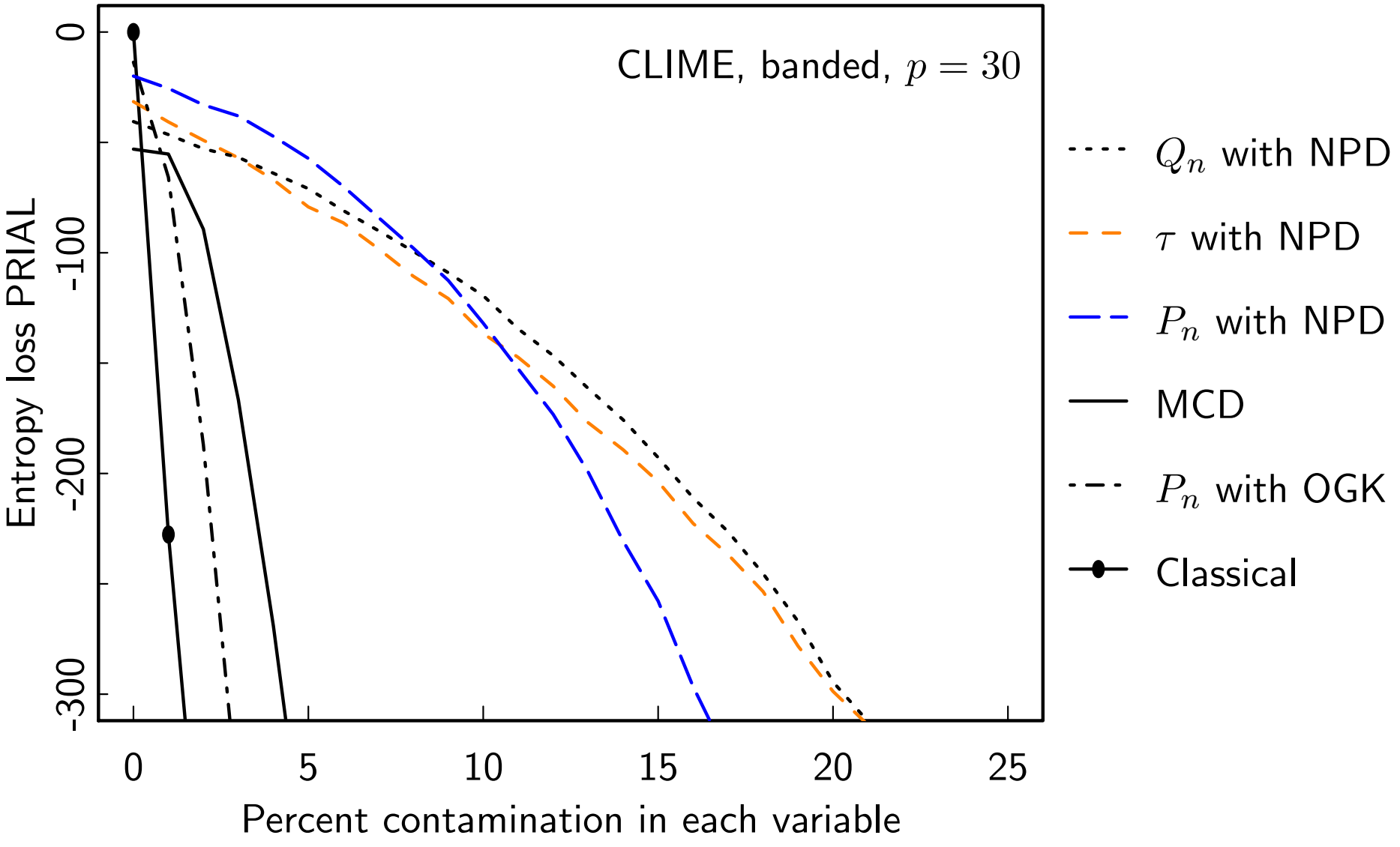
Changing dimension: $p = 90$



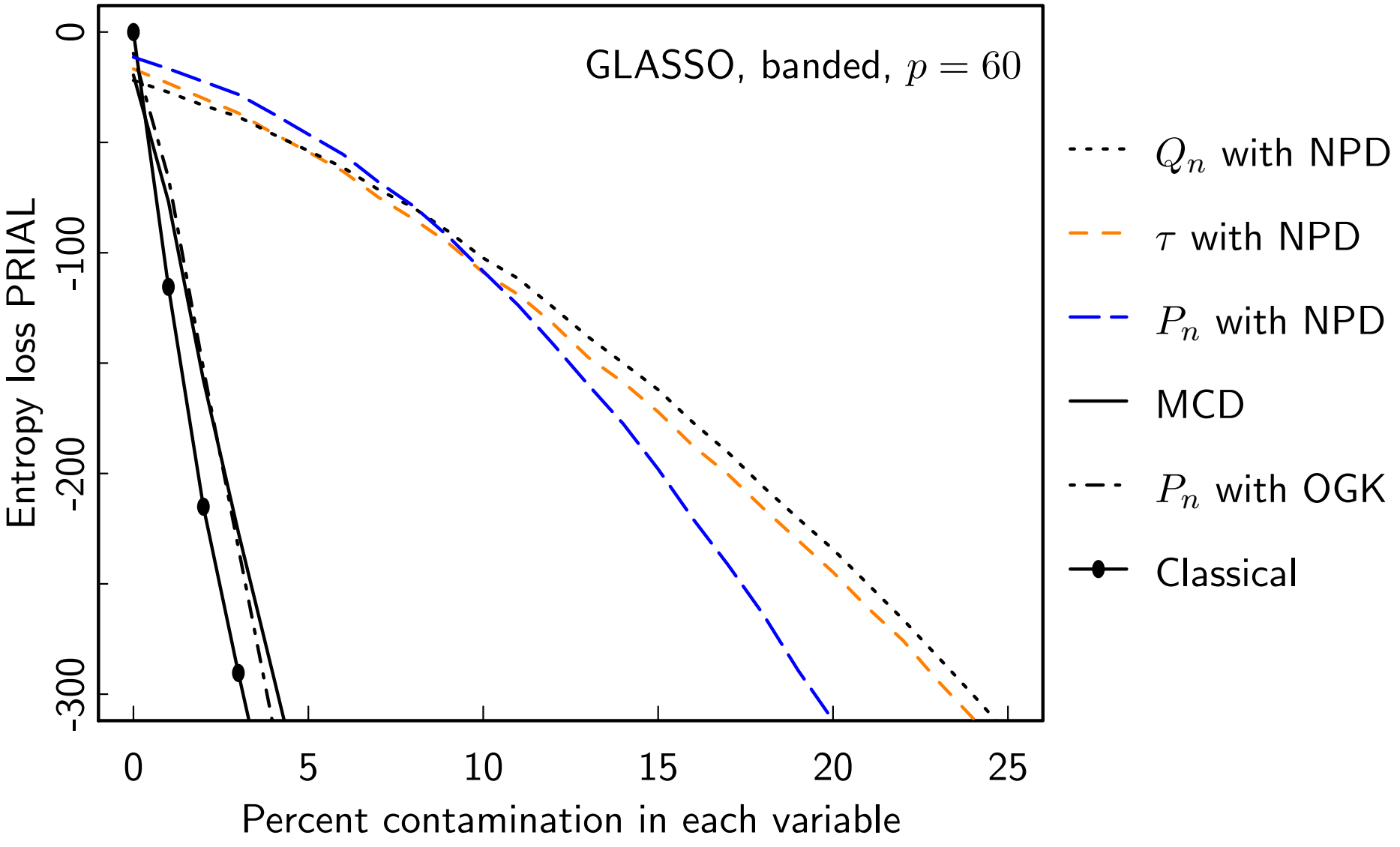
Changing dimension: $p = 60$



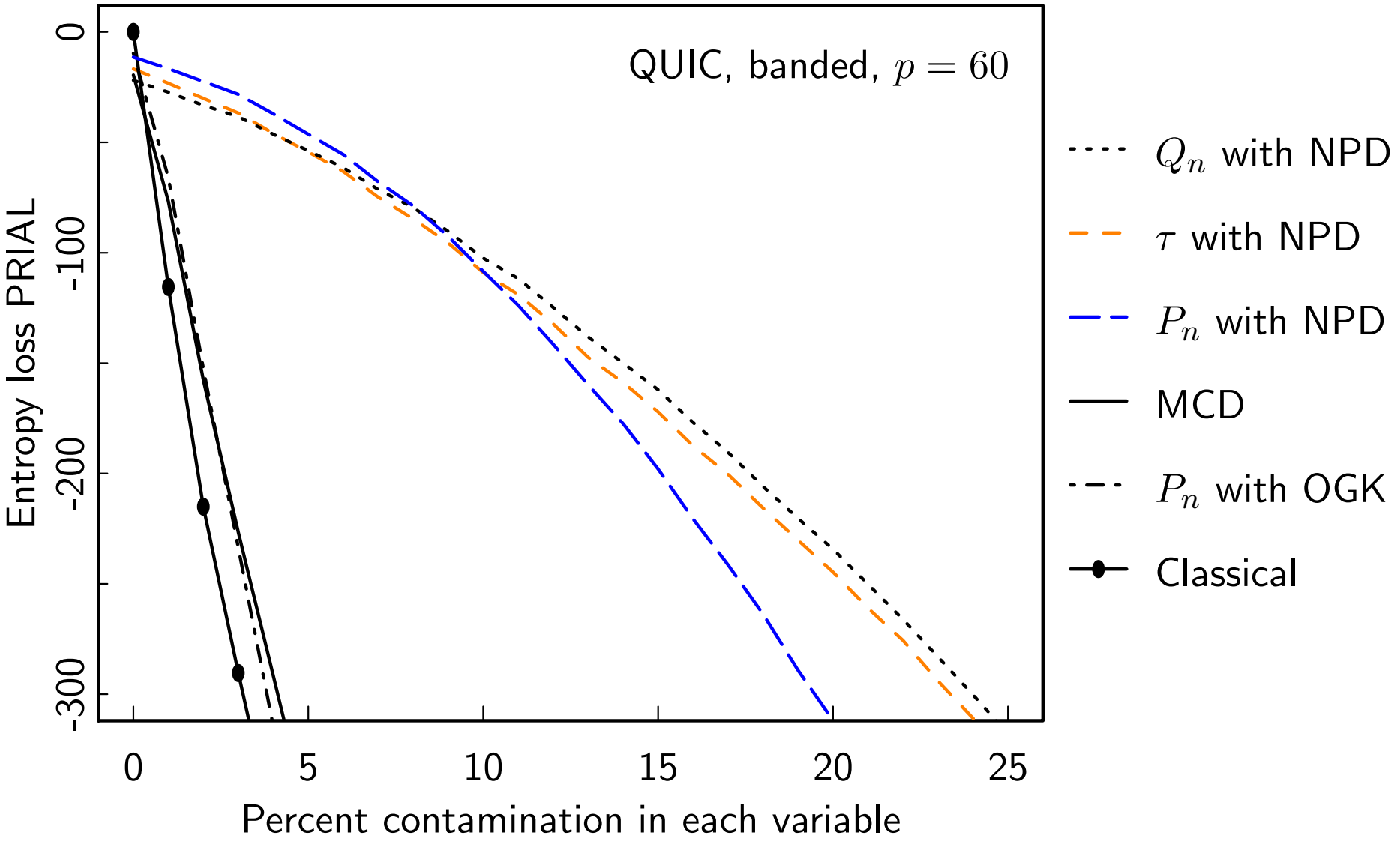
Changing dimension: $p = 30$



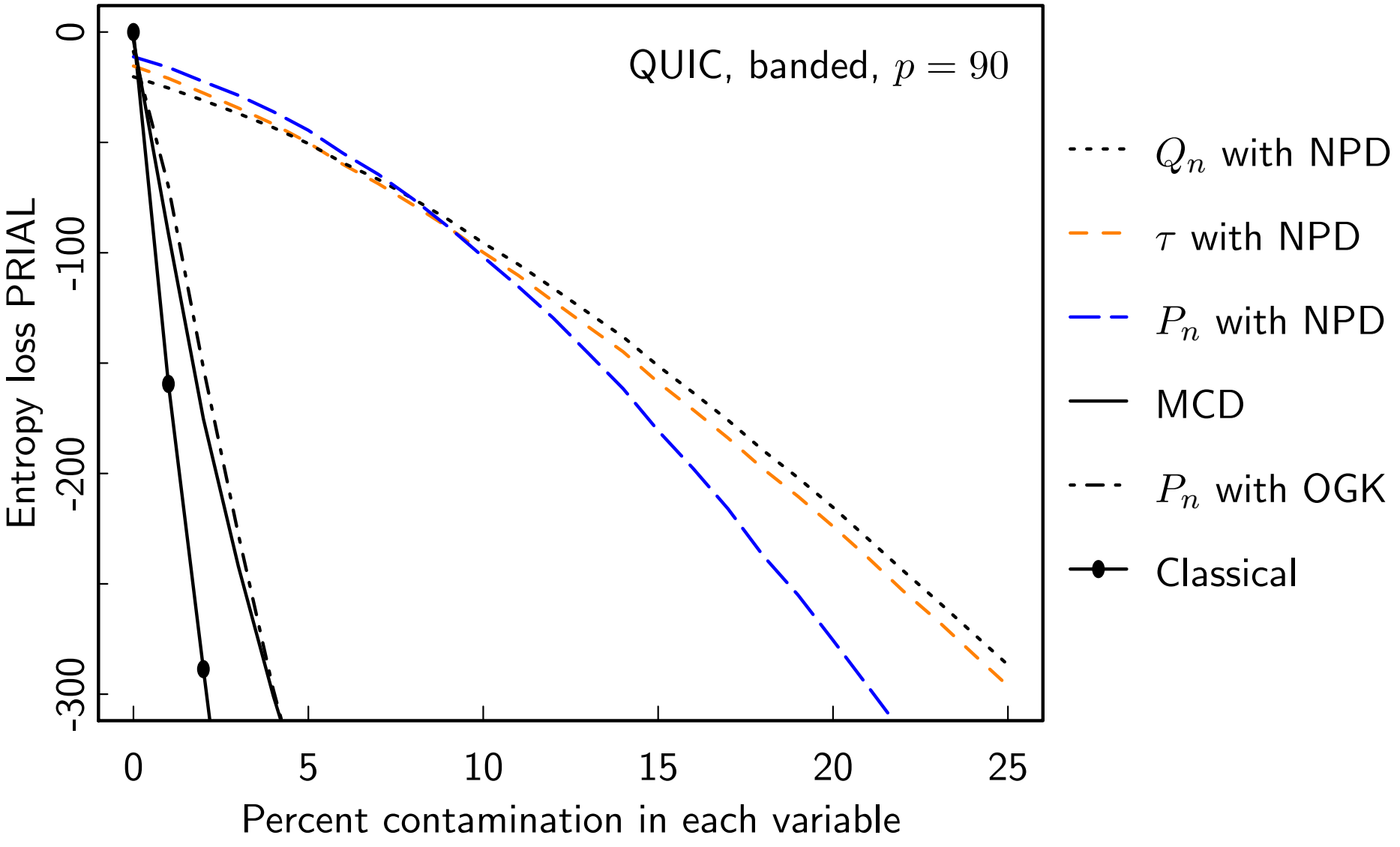
Changing regularisation routine: GLASSO



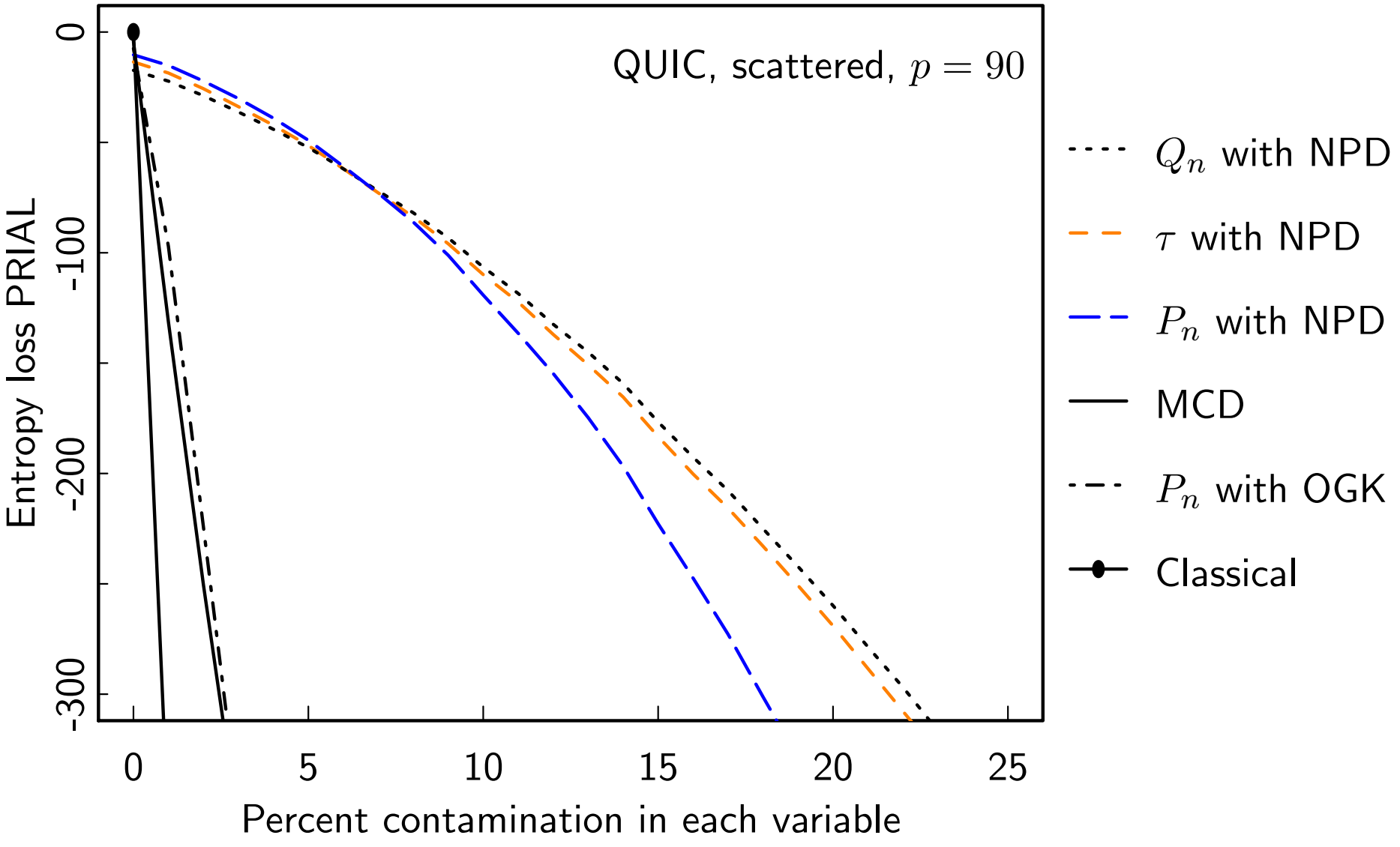
Changing regularisation routine: QUIC



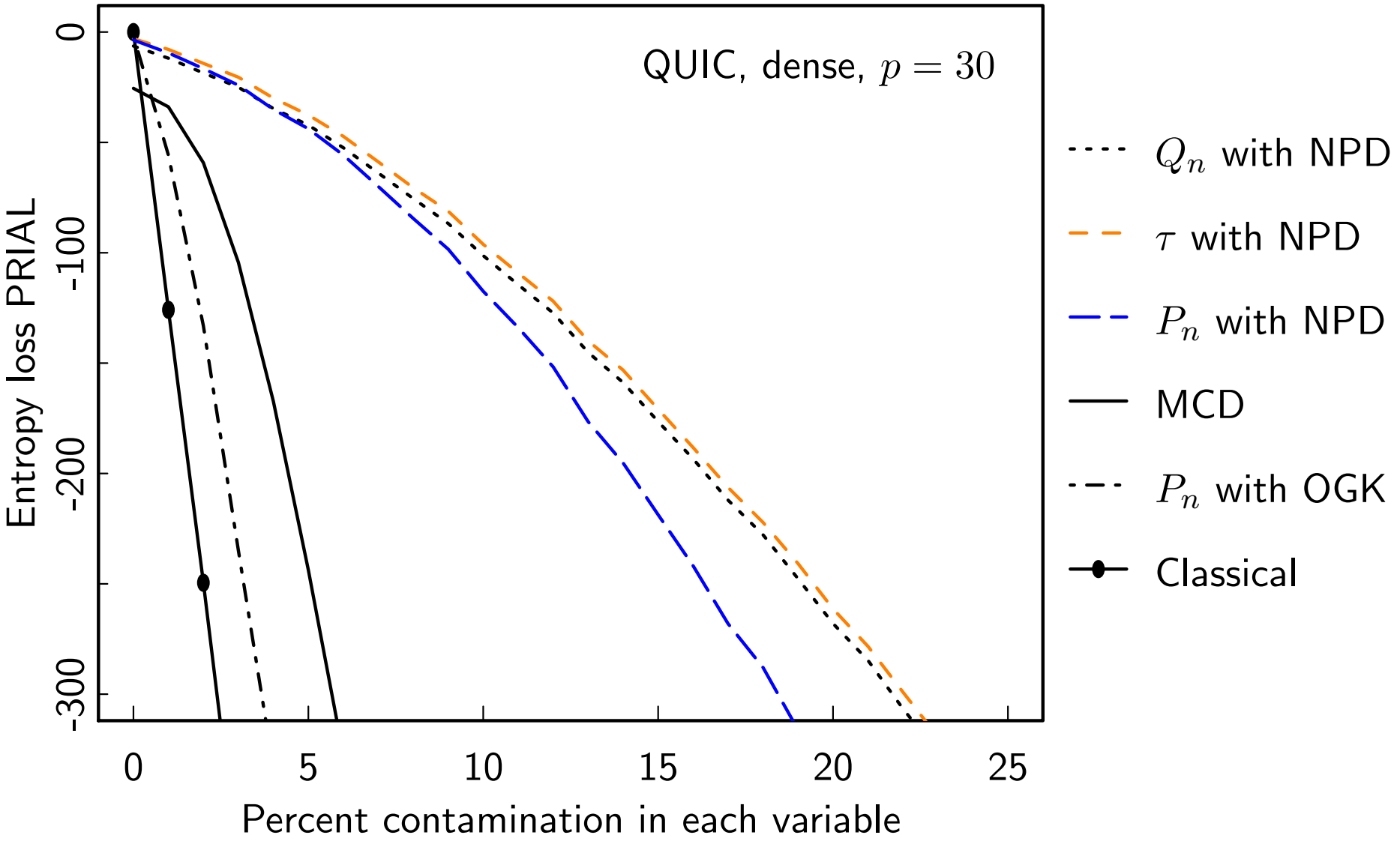
Changing experiment: banded



Changing experiment: sparse



Changing experiment: dense



Estimating dependence

Problem:

To estimate dependence in multivariate settings with cellwise contamination.

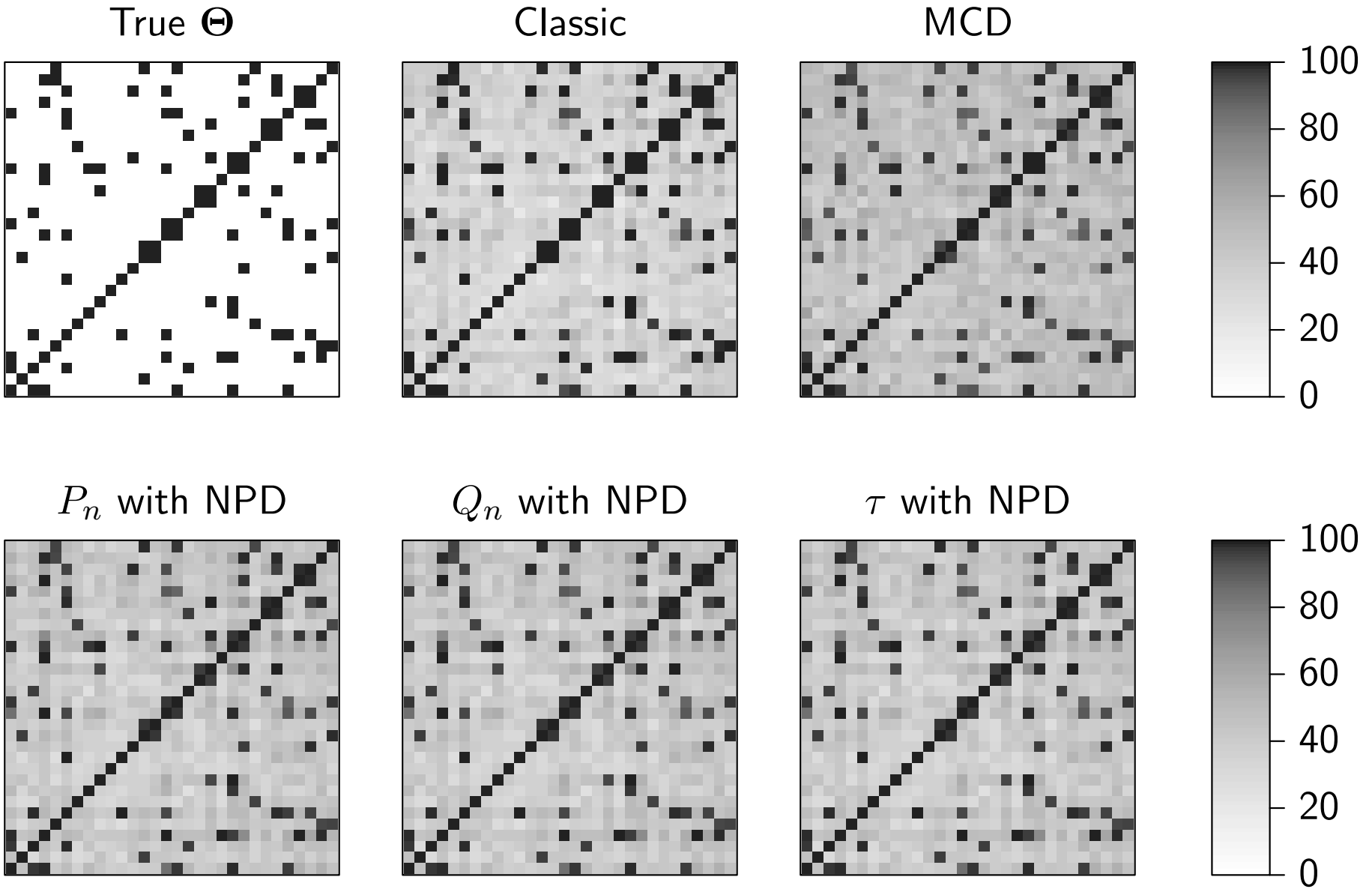
Solution:

1. pairwise covariance matrices
2. correct for positive definiteness
3. regularisation procedure

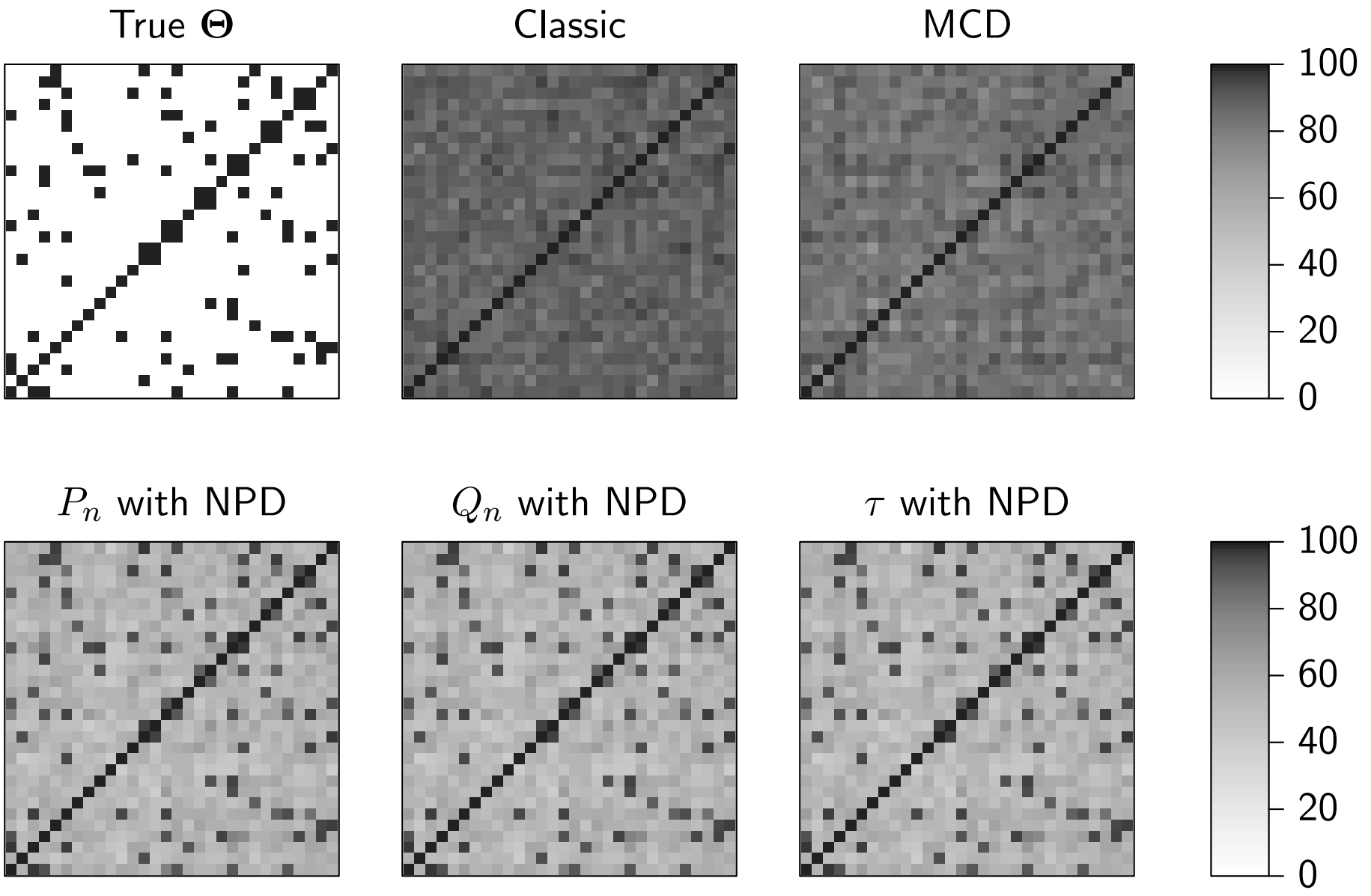
Evaluation:

- Is the estimate "close" to the truth?
- Are we able to recover the support of a Gaussian graphical model?

QUIC, uncontaminated data, $N = 100$ replications



QUIC, 10% contamination, $N = 100$ replications



Evaluating performance

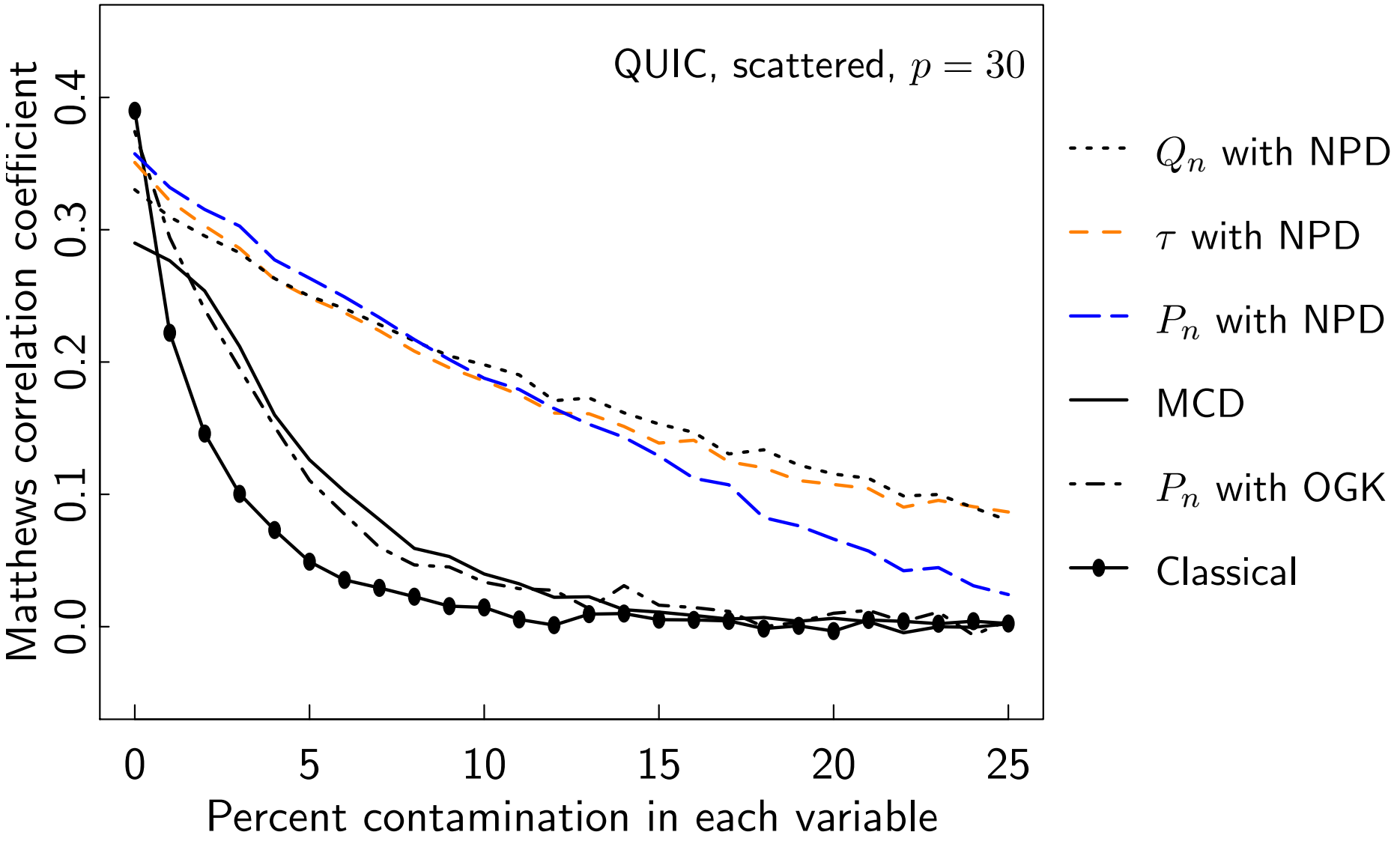
Matthew's Correlation Coefficient (MCC)

Takes into account the number of true positives (TP), false positives (FP), true negatives (TN) and false negatives (FN),

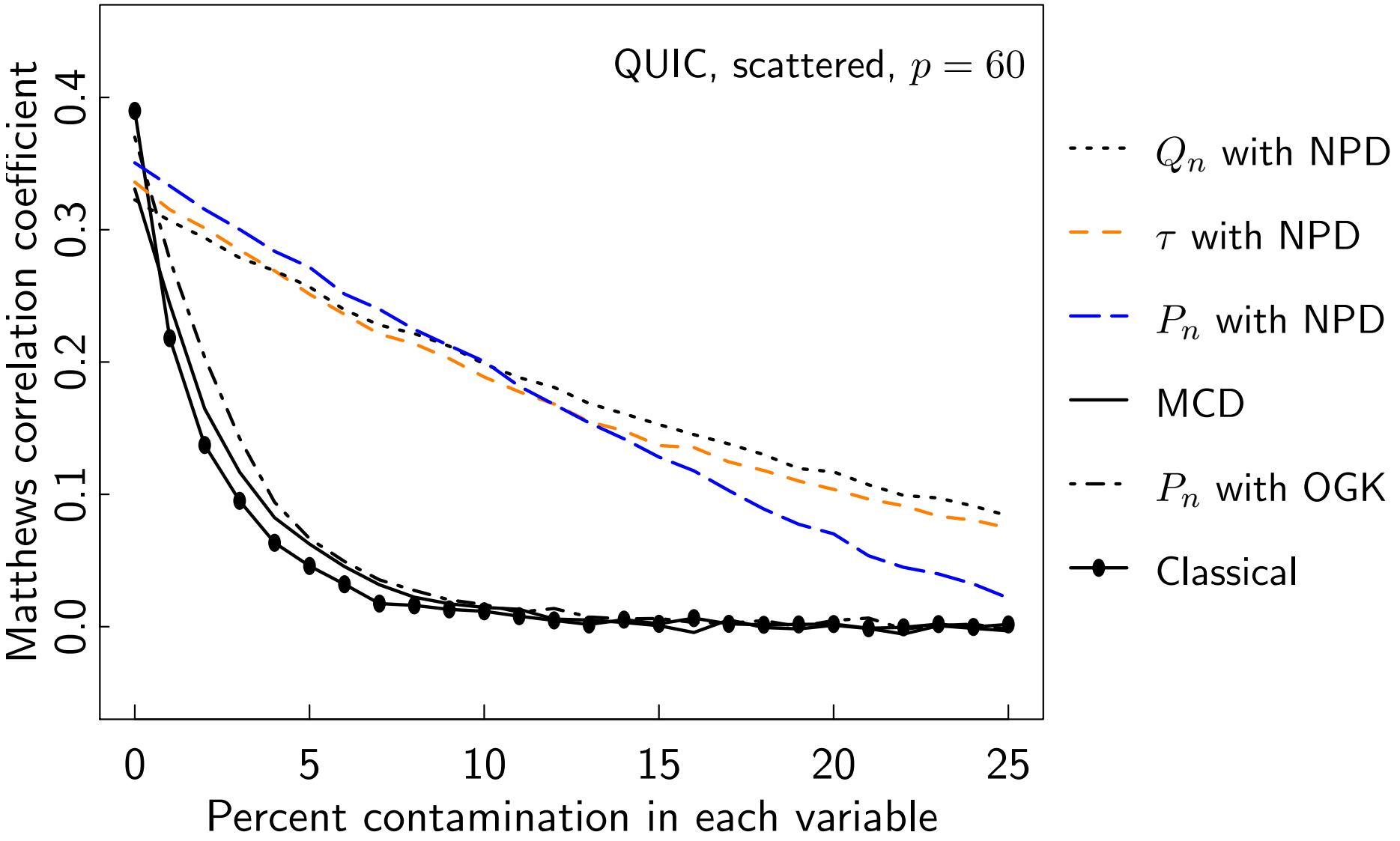
$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN})(\mathrm{TN} + \mathrm{FP})(\mathrm{TN} + \mathrm{FN})}}.$$

Basically the correlation between the observed and predicted binary classifications.

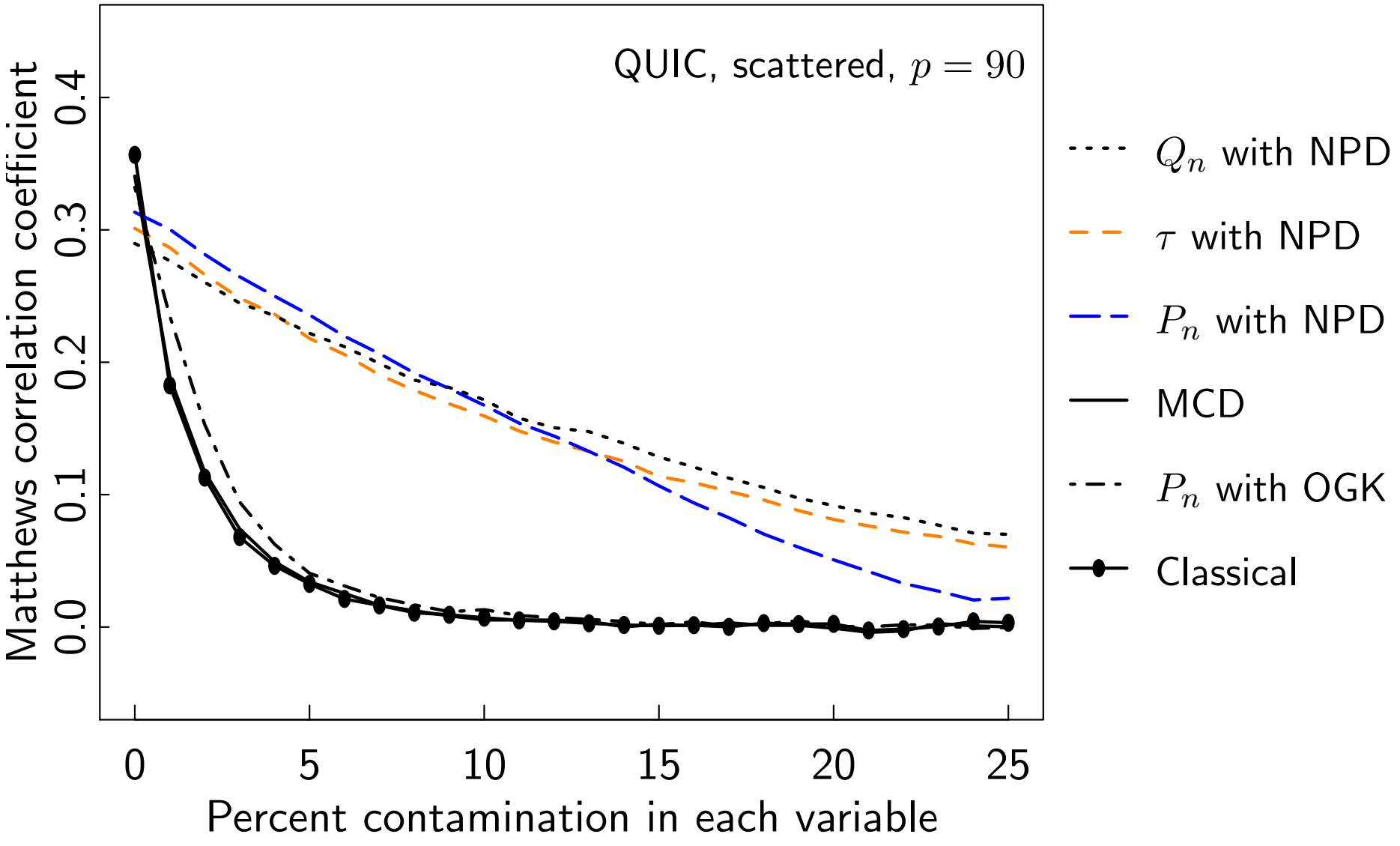
MCC: $p = 30$


QUIC, scattered, $p = 30$
Matthews correlation coefficient
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25
Percent contamination in each variable
Q_n with NPD
τ with NPD
P_n with NPD
MCD
P_n with OGK
Classical

MCC: $p = 60$


QUIC, scattered, $p = 60$
Matthews correlation coefficient
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25
Percent contamination in each variable
Q_n with NPD
τ with NPD
P_n with NPD
MCD
P_n with OGK
Classical

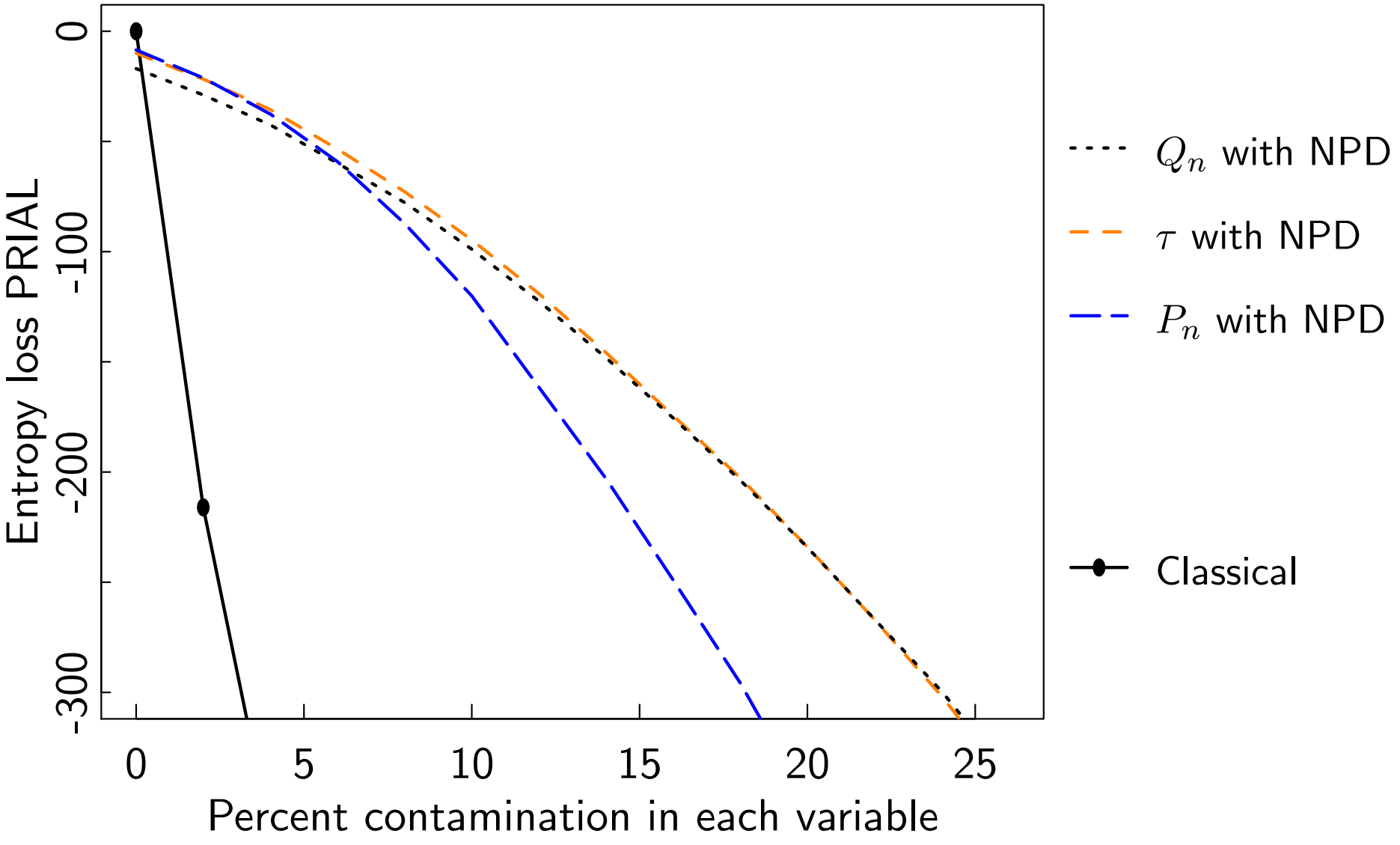
MCC: $p = 90$


QUIC, scattered, $p = 90$
Matthews correlation coefficient
0.0
0.1
0.2
0.3
0.4
0
5
10
15
20
25
Percent contamination in each variable
Q_n with NPD
τ with NPD
P_n with NPD
MCD
P_n with OGK
Classical

Other considerations

- Contaminating model: compare with the missingness literature
- Choice of sparsity parameter: the billion euro question
- $p > n$: looks promising

QUIC, $n = 50$, $p = 60$, sparse precision matrix



Outline

Summary

Problem:

To estimate dependence in multivariate settings with cellwise contamination.

Solution:

1. pairwise covariance matrices
2. correct for positive definiteness
3. regularisation procedure

Result:

- Performs well with moderate amounts of outliers
- Looks promising for $p > n$ problems

References

Alqallaf, F., Van Aelst, S., Yohai, V.J., Zamar, R.H., (2009).

Propagation of outliers in multivariate data.

The Annals of Statistics, 37:311–331.

Cai, T., Liu, W. and Luo, X. (2011).

A constrained ℓ_1 minimization approach to sparse precision matrix estimation.

Journal of the American Statistical Association, 106:594–607.

Friedman, J., Hastie, T. and Tibshirani, R. (2008).

Sparse inverse covariance estimation with the graphical lasso.

Biostatistics, 9(3):432–441.

Gnanadesikan, R. and Kettenring J. R. (1972).

Robust estimates, residuals and outlier detection with multiresponse data.

Biometrics, 28(1):81–124.

References

Higham, N. J. (2002).

Computing the nearest correlation matrix–a problem from finance

IMA Journal of Numerical Analysis, 22(3):329–343.

Hsieh, C-J., Sustik, M.A., Dhillon I.S. and Ravikumar, P.K. (2011).

Sparse inverse covariance matrix estimation using quadratic approximation.

Advances in Neural Information Processing Systems 24, 2330–2338.

Maronna, R. and Zamar, R., (2002).

Robust estimates of location and dispersion for high-dimensional datasets.

Technometrics, 44(4):307–317.

Tarr, G., Müller, S. and Weber, N.C., (2012).

A robust scale estimator based on pairwise means.

Journal of Nonparametric Statistics, 24(1):187–199.