

# On estimating covariances between many assets with histories of highly variable length

Robert B. Gramacy  
Statistical Laboratory  
University of Cambridge  
bobby@statslab.cam.ac.uk

Joo Hee Lee  
Fidelity Investments  
London  
joohee.lee@uk.fid-intl.com

Quantitative portfolio allocation requires the accurate and tractable estimation of covariances between a large number of assets. Asset return histories can greatly vary in length due to the fact that they have started being publicly traded at different times. Such data are said to follow a monotone missingness pattern, under which the likelihood has a convenient factorisation. Upon further assuming that asset returns are multivariate normally distributed, with histories at least as long as the total asset count, maximum likelihood (ML) estimates are easily obtained by performing repeated ordinary least squares (OLS) regressions, one for each asset. Things get more interesting when there are more assets than historical returns. OLS becomes unstable due to rank-deficient design matrices, which is called a “big  $p$  small  $n$ ” problem. We explore remedies that involve making a change of basis, as in principal components or partial least squares regression, or by applying shrinkage methods like ridge regression or the lasso (Hastie et al., 2001). This enables the estimation of covariances between large sets of assets with histories of essentially arbitrary length. Our methods are demonstrated on randomly generated data, and on real financial time series. In essence, we show how the methods of Stambaugh (1996) can be applied for large numbers of assets. Whereas Stambaugh demonstrated his methodology on 22 assets, we show how our approach—essentially the same methodology with a different regression method—can handle thousands. We argue that even when OLS regressions suffice, the more parsimonious ones can offer improvements in both accuracy and interpretation. An accompanying R package called `monomvn` has been made freely available on CRAN (R Development Core Team, 2007).

After formally defining the monotone missingness pattern, we derive the corresponding factorised likelihood, and give the classic algorithm of repeated regressions to analytically find a ML estimator for MVN data. We shall outline methods for dealing with the “big  $p$  small  $n$ ” problem in the context of regression with transformed inputs and shrinkage estimators, highlighting the benefits of increased applicability, accuracy, and interpretability. We will then give the details of a new algorithm for ML estimation of MVN data under the monotone missingness pattern that employs these parsimonious regression algorithms. Our results on synthetic and real (financial) data are benchmarked against several standard comparators, including the Expectation Conditional Maximisation (ECM) algorithm (Meng and Rubin, 1993). We conclude with a discussion that focuses on the ramifications of applying this technique in the context of mean-variance

portfolio rebalancing, and some limitations inherent in taking a maximum likelihood approach.

## References

Hastie, T., Tibshirani, R., and Friedman, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer.

Meng, X. and Rubin, D. B. (1993). “Maximum Likelihood Estimation via the ECM algorithm.” *Biometrika*, 80, 2, 267–278.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Stambaugh, R. F. (1996). “Analyzing Investments Whose Histories Differ in Length.” *Journal of Financial Economics*, 45, 285–331.