Errata from The Book of Alternative Data

By Alexander Denev and Saeed Amen

Updated 08 December 2020

Page 13

Figure 1.3 – caption should read Q1 2011-Q3 2019; correlation 47%

Missing References in Bibliography

Agarwal, R., Imielinksi, T. and Swami, A. (1993). *Mining association rules between sets of items in large databases*. Proceedings of the ACM SIGMOD, 207-16

Ahelegbey, D.F. and Giudici, P., (2014). *Hierarchical graphical models with application to systemic risk*. *University Ca'Foscari of Venice, Dept. of Economics Research Paper Series No*, *1*.

Andersen, T.G., Bollerslev, T., Diebold, F.X. and Vega, C., (2007). *Real-time price discovery in global stock, bond and foreign exchange markets*. *Journal of international Economics*, *73*(2), pp.251-277.

Ang, A., (2014). *Asset management: A systematic approach to factor investing*. Oxford University Press.

Angiulli, F. & Fassetti, F. & Manco, G. & Palopoli, L. (2017). *Outlying property detection with numerical attributes*. Data Min. Knowl. Discov. 31(1): 134-163

Angiulli, F. & Fassetti, F. & Palopoli, L. (2009). *Detecting Outlying Properties of Exceptional Objects*. ACM Trans. Database Syst. 34. 10.1145/1508857.1508864.

Authers, J. (2015). *Why factor investing is flavour of the month*. Financial Times. 28 Jan 2015

Back, K. (2010). *Asset pricing and portfolio choice theory*. Oxford University Press.

Barnard, J. & Rubin, D. (1999). *Small-sample degrees of freedom with multiple imputation*. Biometrika. 86. 948. 10.1093/biomet/86.4.948.

Barnett, V. & Lewis, T. (1978). *Outliers in Statistical Data*. First Edition. John Wiley & Sons.

Bauer, J., Angelini, O. and Denev, A., (2017). Imputation of multivariate time series data-performance benchmarks for multiple imputation and spectral techniques.

Beckers, J.M. and Rixen, M., (2003). *EOF calculations and data filling from incomplete oceanographic datasets*. *Journal of Atmospheric and oceanic technology*, *20*(12), pp.1839-1856.

Billio, M., Caporin, M., Panzica, R. and Pelizzon, L., (2016). *The impact of network connectivity on factor exposures, asset pricing and portfolio diversification*.

Bollen, J., Mao, H. and Zeng, X., (2011). *Twitter mood predicts the stock market*. *Journal of computational science*, *2*(1), pp.1-8.

Borovkova, Svetlana & Lammers, Philip. (2017). *Sector News Sentiment Indices*. 10.13140/RG.2.2.14691.25125.

Breiman, L., (2001). Random forests. *Machine learning*, *45*(1), pp.5-32.

Breunig, Markus & Kriegel, Hans-Peter & Ng, Raymond & Sander, Joerg. (2000). *LOF: Identifying Density-Based Local Outliers*. ACM Sigmod Record. 29. 93-104. 10.1145/342009.335388.

Buuren, S.V. and Groothuis-Oudshoorn, K., (2010). *mice: Multivariate imputation by chained equations in R*. *Journal of statistical software*, pp.1-68.

Carhart, M.M., (1997). *On persistence in mutual fund performance*. *The Journal of Finance*, *52*(1), pp.57-82.

Chandola, V., Banerjee, A. and Kumar, V., (2009). *Anomaly detection: A survey*. *ACM computing surveys (CSUR)*, *41*(3), p.15.

Chapados, N. and Bengio, Y., (2007), June. *Forecasting and Trading Commodity Contract Spreads with Gaussian Processes*. In *13th International Conference on Computing in Economics and Finance*.

Christen, P., (2012). *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media.

Clarke, R.G., de Silva, H. and Murdock, R., (2005). *A factor approach to asset allocation*. *Journal of Portfolio Management*, *32*(1), p.10.

Cochrane, J.H., (2009). *Asset pricing: Revised edition*. Princeton university press.

Connor, G., (1995). The three types of factor models: A comparison of their explanatory power. *Financial Analysts Journal*, *51*(3), pp.42-46.

Connor, G., Goldberg, L.R. and Korajczyk, R.A., (2010). *Portfolio risk analysis*. Princeton University Press.

Corey, D.M., Dunlap, W.P. and Burke, M.J., (1998). *Averaging correlations: Expected values and bias in combined Pearson rs and Fisher's z transformations*. *The Journal of General Psychology*, *125*(3), pp.245-261.

De Prado, M.L., (2018). *Advances in financial machine learning*. John Wiley & Sons.

Dixon, M. and Halperin, I., & Bilokon, P. (2020). *Machine learning in finance: From Theory to Practice*. Springer.

Duan, L. & Tang, G. & Pei, J. & Bailey, J. & Campbell, A. & Tang, C. (2015). *Mining outlying aspects on numeric data. Data Mining and Knowledge Discovery*. 29. 10.1007/s10618-014-0398-2.

Enders, C.K. (2010). *Applied missing data analysis*. Guilford press.

Ester, M.; Kriegel, H.-P.; Sander, J. & Xu, X. (1996). *A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*, *in* Evangelos Simoudis; Jiawei Han & Usama M. Fayyad, ed., 'KDD' , AAAI Press, , pp. 226-231 .

Fama, E.F. and French, K.R., (1992). *The cross-section of expected stock returns*. *The Journal of Finance*, *47*(2), pp.427-465.

Fama, E.F. and French, K.R. (1993). *Common risk factors in the returns on stocks and bonds*. *Journal of financial economics*, *33*(1), pp.3-56.

Fama, E.F. and French, K.R. (1995). *Size and book-to-market factors in earnings and returns*. *The Journal of Finance*, *50*(1), pp.131-155.

Fama, E.F. and French, K.R. (1996). *Multifactor explanations of asset pricing anomalies*. *The Journal of Finance*, *51*(1), pp.55-84.

Fama, E.F. and French, K.R. (2004). *The capital asset pricing model: Theory and evidence*. *Journal of economic perspectives*, *18*(3), pp.25-46.

Farhangfar, A. & Kurgan, L. & Dy, J. (2008). *Impact of imputation of missing values on classification error for discrete data*. Pattern Recognition. 41. 3692-3705. 10.1016/j.patcog.2008.05.019.

Fawcett, T. (2006). *An Introduction to ROC Analysis*. Pattern Recognition Letters, 27, 861--874. doi: 10.1016/j.patrec.2005.10.010

Financial Times (2017, 24 Feb). *Insurance and the big data technology revolution*.

García Laencina, P. & Sancho-Gómez, J. L. & Figueiras-Vidal, A. (2010). *Pattern classification with missing data: A review*. Neural Computing and Applications. 19. 263-282. 10.1007/s00521-009-0295-6.

Geirhos, R., Janssen, D.H., Schütt, H.H., Rauber, J., Bethge, M. and Wichmann, F.A., (2017). *Comparing deep neural networks against humans: object recognition when the signal gets weaker*. *arXiv preprint arXiv:1706.06969*.

Goldstein, M. (2014). *Anomaly Detection in Large Datasets*.

Goldstein, M. & Dengel, A. (2012). *Histogram-based Outlier Score (HBOS): A fast Unsupervised Anomaly Detection Algorithm*.

Goldstein, M. and Uchida, S. (2016) *A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data*. PLoS ONE 11(4): e0152173. https://doi.org/10.1371/journal.pone.0152173.

Golyandina, N., Korobeynikov, A., Shlemov, A. and Usevich, K., (2013). *Multivariate and 2D extensions of singular spectrum analysis with the Rssa package*. *arXiv preprint arXiv:1309.5050*.

Gomes, P. and Peraita, E.V., (2016). *The Effects of Announcements of Leading and Sentiments Indicators on Euro Area Financial Markets*.

Graham, J. (2009). *Missing Data Analysis: Making It Work in the Real World*. Annual review of psychology. 60. 549-76. 10.1146/annurev.psych.58.110405.085530

Grzymala-Busse, J.W., and Hu, M. (2000). *A Comparison of Several Approaches to Missing Attribute Values in Data Mining*.

Han, J., Kamber, M. and Pei, J. (2011). *Data Mining: Concepts and Techniques*. 3rd Edition, Morgan Kaufmann Publishers, Burlington.

Hanousek, J. and Kočenda, E., (2011). Foreign news and spillovers in emerging European stock markets. *Review of International Economics*, *19*(1), pp.170-188.

Hawkins D. (1980). *Identification of Outliers. Chapman and Hall*.

Heckman, J.R., Boehmer, E.L., Peters, E.H., Davaloo, M. and Kurup, N.G., (2015). *A pricing model for data markets*. *iConference 2015 Proceedings*.

Hess, D., Huang, H. and Niessen, A., (2008). *How do commodity futures respond to macroeconomic news? Financial Markets and Portfolio Management*, *22*(2), pp.127-146.

Honaker, J., King, G. and Blackwell, M., (2011). *Amelia II: A program for missing data*. *Journal of statistical software*, *45*(7), pp.1-47.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R, 1st Edition*. Springer-Verlag, New York.

Jerez, J., Molina, I., García Laencina, P., Alba, E., Ribelles, N., Martin, M. & Franco, L. (2010). *Missing Data Imputation Using Statistical and Machine Learning Methods in a Real Breast Cancer Problem*. Artificial Intelligence in Medicine, 50, 105-115. Artificial intelligence in medicine. 50. 105-15. 10.1016/j.artmed.2010.05.002.

Johnson, M.A. and Watson, K.J., (2011). *Can Changes in the Purchasing Managers' Index Foretell Stock Returns? An Additional Forward-Looking Sentiment Indicator*. *The Journal of Investing*, *20*(4), pp.89-98.

Jones, C. I. & Tonetti, C. (2019). *Nonrivalry and the Economics of Data*. No. w26260. National Bureau of Economic Research, 2019

Kang, P. (2013). *Locally linear reconstruction based missing value imputation for supervised learning*. Neurocomputing. 118. 65-78. 10.1016/j.neucom.2013.02.016.

Kaufman, L. and Rousseeuw, P. J. (2008). *In Finding Groups in Data*. John Wiley and Sons, Inc.

Knorr, E. & Ng, R. (1996). *Finding Aggregate Proximity Relationships and Commonalities in Spatial Data Mining*. IEEE Trans. Knowl. Data Eng.. 8. 884-897. 10.1109/69.553156.

Kofman, P. and Sharpe, I.G., (2003). *Using multiple imputation in the analysis of incomplete observations in finance*. *Journal of Financial Econometrics*, *1*(2), pp.216-249.

Koller, D., Friedman, N. and Bach, F., (2009). *Probabilistic graphical models: principles and techniques*. MIT press.

Kondrashov, D. and Ghil, M., (2006). *Spatio-temporal filling of missing points in geophysical data sets*. *Nonlinear Processes in Geophysics*, *13*(2), pp.151-159.

Kriegel, H-P. & Schubert, M. & Zimek, A. (2008). *Angle-based outlier detection in high-dimensional data*. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 444-452. 10.1145/1401890.1401946.

Laney, D.B., (2017). *Infonomics: how to monetize, manage, and measure information as an asset for competitive advantage*. Routledge

Lintner, J., (1965). *Security prices, risk, and maximal gains from diversification*. *The Journal of Finance*, *20*(4), pp.587-615.

Little, R.J. and Rubin, D.B., (2019). *Statistical analysis with missing data* (Vol. 793). Wiley.

Little, R.J., (1988). *A test of missing completely at random for multivariate data with missing values*. *Journal of the American statistical Association*, *83*(404), pp.1198-1202.

Liu, Fei Tony & Ting, Kai & Zhou, Zhi-Hua. (2012). *Isolation-Based Anomaly Detection. ACM Transactions on Knowledge Discovery From Data* - TKDD. 6. 1-39. 10.1145/2133360.2133363.

Löser, A. & Stahl, F. & Muschalle, A. & Vossen, G. (2012). *Pricing Approaches for Data Markets*. 10.1007/978-3-642-39872-8_10.

Luengo, J. & García, S.& Herrera, F. (2011). *On the choice of the best imputation methods for missing values considering three groups of classification methods*. Knowledge and Information Systems - KAIS. 32. 1-32. 10.1007/s10115-011-0424-2.

Marenzi, O., (2017). *Alternative Data – The New Frontier in Asset Management*. Opimas, March 31, 2017

Markowitz, H., (1991). *Portfolio selection: efficient diversification of investments*. Blackwell

Markowitz, H.M. and Todd, G.P., (2000). *Mean-variance analysis in portfolio choice and capital markets* (Vol. 66). John Wiley & Sons.

McKinsey (2016). *McKinsey Quarterly, Straight talk about big data*. October 2016

Micenkova, B., Ng, R., Dang, X-H. & Assent, I. (2013). *Explaining Outliers by Subspace Separability*. Proceedings - IEEE International Conference on Data Mining, ICDM. 518-527. 10.1109/ICDM.2013.132.

Miller, G., (2006). *Needles, haystacks, and hidden factors*. *Journal of Portfolio Management*, *32*(2), p.25.

Mok, A., and Saha, R., (2017). *"Strategic risk management in banking," Part 1*, Inside Magazine – Edition 2017, Deloitte

Mossin, J., (1966). *Equilibrium in a capital asset market*. *Econometrica: Journal of the econometric society*, pp.768-783.

Ng, E. quotation from article by Blakeslee, Sandra (1990). *Lost on Earth: Wealth of Data Found in Space*. The New York Times.

Neudata (2020). *Alternative data intelligence*

Ramaswamy, S., Rastogi, R., Shim, K. & Korea, T. (2000). *Efficient Algorithms for Mining Outliers from Large Data Sets*. 10.1145/342009.335437.

Rätsch, G., Schölkopf, B., Mika, S. & Müller, K-R. (2000). *SVM and boosting: One class*.

Rezvan, P.H., Lee, K.J. and Simpson, J.A., (2015). *The rise of multiple imputation: a review of the reporting and implementation of the method in medical research*. *BMC medical research methodology*, *15*(1), p.30.

Risk, (2019). Quants say big data is all buzz, no alpha

Rosenblatt, M. (1956). *"Remarks on Some Nonparametric Estimates of a Density Function".* The Annals of Mathematical Statistics. 27 (3): 832–837. doi:10.1214/aoms/1177728190.

Ross, S.A., (1972). *Portfolio and Capital Market Theory with Arbitrary Preferences and Distributions: The General Validity of the Mean-Variance Approach in Large Markets* (No. 12-72). Wharton School Rodney L. White Center for Financial Research

Ross, S.A., (1973). *Return, risk and arbitrage*. Rodney L. White Center for Financial Research, The Wharton School, University of Pennyslvania.

Ross, S.A., (2013). *The arbitrage theory of capital asset pricing*. In *Handbook of the fundamentals of financial decision making: Part I* (pp. 11-30).

Schafer, J.L., (1997). *Analysis of incomplete multivariate data*. Chapman and Hall/CRC.

Schaffer, C., (1994). *A conservation law for generalization performance*. In *Machine Learning Proceedings 1994* (pp. 259-265). Morgan Kaufmann.

Schölkopf, B., Platt, J., Shawe-Taylor, J., Smola, A. and Williamson, R. (2001). *Estimating the Support of a High-Dimensional Distribution*. Neural Comput. 13, 7 (July 2001), 1443–1471. DOI:https://doi.org/10.1162/089976601750264965.

Sharpe, W.F., (1964). *Capital asset prices: A theory of market equilibrium under conditions of risk*. *The Journal of Finance*, *19*(3), pp.425-442.

Short, J. and Todd, S., (2017). What's Your Data Worth?. *MIT Sloan Management Review*, *58*(3), p.17

Soe, A. M. and Poirier, R. (2016). *SPIVA US Year-End 2016 Scorecard*. Retrieved from: https://us.spindices.com/documents/spiva/spiva-us-year-end-2016.pdf

Stekhoven, D.J. and Bühlmann, P., (2011). *MissForest - non-parametric missing value imputation for mixed-type data*. *Bioinformatics*, *28*(1), pp.112-118.

Su, Y.S., Gelman, A.E., Hill, J. and Yajima, M., (2011). *Multiple imputation with diagnostics (mi) in R: Opening windows into the black box*.

Sugiyama, M. and Kawanabe, M., (2012). *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press.

Tan, P.N., Steinbach, M. and Kumar, V. (2006). *Introduction to Data Mining*. Pearson: Addison Wesley, Boston.

Treynor, J.L., (1962). *Jack Treynor's' Toward a Theory of Market Value of Risky Assets'. Available at SSRN 628187*.

Turner, M. A., (2011). *Give Credit where Credit is Due*. Political and Economic Research Council.

Turner, M. A., Lee, A., Varghese, R., Walker, P., (2008). *You Score You Win*. Political and Economic Research Council.

United Nations, (2015). *Global Working Group on Big Data for Official Statistics Task Team on Cross-Cutting Issues, Deliverable 2: Revision and Further Development of the Classification of Big Data*

Vinh, N., Chan, J., Romano, S., Bailey, J., Leckie, C., Ramamohanarao, K., & Pei, J. (2016). *Discovering outlying aspects in large datasets*. Data Mining and Knowledge Discovery. 30. 10.1007/s10618-016-0453-2.

Wang, Hai & Wang, Shouhong. (2010). *Mining incomplete survey data through classification*. Knowl. Inf. Syst.. 24. 221-233. 10.1007/s10115-009-0245-8.

Witten, I.H., Frank, E. and Hall, M.A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd Edition, Morgan Kaufmann Publishers, Burlington.

Wolpert, D.H., (2002). *The supervised learning no-free-lunch theorems*. In *Soft computing and industry* (pp. 25-42). Springer, London.

Yan, X. and Zheng, L. (2017). *Fundamental analysis and the cross-section of stock returns: A data-mining approach*. The Review of Financial Studies, 30(4):1382-1423.

Yu, H. and Zhang, M., (2017). *Data pricing strategy based on data quality*. *Computers & Industrial Engineering*, *112*, pp.1-10.

Zou, Y., An, A. & Huang, X. (2005). *Evaluation and automatic selection of methods for handling missing data*. 728 - 733 Vol. 2. 10.1109/GRC.2005.1547387.