Preliminary Results

Many applications of modern statistics involve a large number of measurements and can be considered in a linear algebra framework. In many of these problems, the dimensionality p exceeds the number of observations, $N = n + 1$; e.g. DNA Microarray data, portfolio selection in economics, network problems in computer science. Classical multivariate inference techniques become degenerate when $p > n$ and many perform poorly when p is of the same order as n. My early results concentrated on inference of the covariance matrix in large-dimensional situations. This methodology was extended into time series goodness-of-fit testing and has lead to my current research.

High-dimensional Multivariate Analysis

Consider a random sample of p-dimensional independent observations, typically from a multivariate normal distribution $N_p(\mu, \Sigma)$, where both parameters are unknown. For hypothesis testing on Σ, typically a statistic based on the likelihood ratio criterion (LRT) would be used. However, when $p > n$, the LRT is not available, and when p is close to the size of n, the LRT is ill-conditioned since the smallest eigenvalues of the sample covariance matrix will tend towards zero. Statistics were developed for the *sphericity* hypothesis, $\Sigma = \sigma^2 I$, where σ^2 is an unknown scalar proportion in Fisher et al. (2010), and the *identity* hypothesis, $\Sigma = I$, in Fisher (2012). The statistics are based on parametric functions of the first four arithmetic means of the eigenvalues of Σ; i.e. $a_i = (1/p) \sum \lambda_j^i$. The theoretical results in Fisher et al. (2010) and Fisher (2012) show estimators for the first four arithmetic means are asymptotically normally distributed as $(n, p) \to \infty$, under the assumptions that the data is normal and $p/n \to c$, known as a concentration.

In Fisher and Sun (2011), the estimators for the first four arithmetic means were utilized in the application of estimating the covariance matrix. In particular, a new set of estimators were introduced for the optimal intensity of a Stein-type shrinkage estimator (see Stein (1956), Ledoit and Wolf (2003)): $S^* = \lambda F + (1 - \lambda)S$, where S is the sample covariance matrix, F is a well-conditioned target matrix and $\lambda \in [0, 1]$ is known as the shrinkage intensity. When n is large compared to p, $\lambda \to 0$, and as p increases to (or exceeds) n, more weight is put on the target matrix. An optimal intensity with respect to the squared loss function based on the Frobenius norm can be found and we suggest (n, p)-consistent estimators based on the work of Srivastava (2005) for three classic targets: the identity matrix, a diagonal with a common variance, and the diagonal matrix consisting of the diagonal elements of the sample covariance matrix.

Time Series Diagnostic Testing

Many time series data are affected by serial-correlation. Empirical evidence in economics suggest some series may suffer from changes in variability, a phenomenon known as heteroscedasticity. To model the serial correlation structure of a time series we typically use a stationary and invertible ARMA(p, q) process where p is the autoregressive order, q is the moving average order and the error (innovation) sequence is iid with zero mean and finite variance. After a time series has been fit with some ARMA model the correlation structure of the residual innovation sequence would be estimated by the sample autocorrelation function

$$\hat{r}_k = \frac{\sum_{k+1}^{n} \hat{\epsilon}_t \hat{\epsilon}_{t-k}}{\sum_{t=1}^{n} \hat{\epsilon}_t^2} \quad \text{for } k = 1, 2, \ldots, m. \tag{1}$$

Here $\{\hat{\epsilon}_t\}$ are the observed residuals after fitting an ARMA(p, q) model. If the fitted model is adequate, each of $\{\hat{r}_k\}$ should be approximately equal to zero where m is chosen large enough to detect meaningful correlation in the data. However, if the fit underestimates the ARMA orders, the values of the autocorrelations should significantly deviate from zero. A diagnostic test for the adequacy of a fitted ARMA model was introduced by Box and Pierce (1970) where they derive the asymptotic joint-distribution of the sample autocorrelations. Ljung and Box (1978) improve on the finite sample performance by standardizing the values and Monti (1994) proposed a test utilizing the standardized residual partial autocorrelations. All three statistics are asymptotically distributed as a chi square random variable with $m-(p+q)$ degrees of freedom when an ARMA(p, q) was fit to the data and have the same general form

$$\widetilde{Q} = n \sum_{k=1}^{m} \frac{n+2}{n-k} \hat{r}_k^2, \tag{2}$$

where $\widetilde{Q}$ is the widely used Ljung and Box (1978) statistic. Simulations in Monti (1994) demonstrate her statistic is more powerful than $\widetilde{Q}$ when the fitted model underestimates the order of the moving average component. Practitioners of statistics typically use the statistic $\widetilde{Q}$.

Recently, Peña and Rodríguez (2002, 2006) and Mahdi and McLeod (2012) proposed statistics based on the likelihood ratio criterion (i.e. determinant) on the estimated Toeplitz matrix of the autocorrelation function:

$$\widehat{\mathbf{R}}_m = \begin{bmatrix} 1 & \hat{r}_1 & \cdots & \hat{r}_m \\ \hat{r}_1 & 1 & \cdots & \hat{r}_{m-1} \\ \vdots & \cdots & \ddots & \vdots \\ \hat{r}_m & \cdots & \hat{r}_1 & 1 \end{bmatrix}. \tag{3}$$

Under the null hypothesis that the model has been correctly identified, the matrix $\widehat{\mathbf{R}}_m$ is close to the identity matrix. Simulation experiments demonstrate these tests can improve over the traditionally used Ljung Box statistic in terms of power. However, there are computational concerns when the matrix $\widehat{\mathbf{R}}_m$ is constructed using standardized autocorrelations. Furthermore, as m-increases, the behavior of $\widehat{\mathbf{R}}_m$ is similar to that in high-dimensional multivariate analysis and the statistics become unstable.

Using the methodology from Srivastava (2005) and my work in high-dimensional analysis, a statistic is constructed for testing the adequacy of a fitted ARMA process based on the trace of the square of $\widehat{\mathbf{R}}_m$:

$$\widetilde{Q}_W = n(n+2) \sum_{k=1}^{m} \frac{(m-k+1)}{m} \frac{\hat{r}_k^2}{n-k}. \tag{4}$$

The statistic $\widetilde{Q}_W$ can be interpreted as a weighted Ljung-Box test. The residual at lag one is given the most weight, 1, while the residual at lag m is given the least weight, $1/m$. The statistic is easy to implement and is computationally stable. A similar derivation using the matrix of partial autocorrelations leads to a weighted Monti statistic.

The proposed weighted portmanteau statistics are quite versatile and can be modified to detect the presence of a nonlinear or heteroscedastic innovation sequence, such as the Generalized Autoregressive Conditional Heteroscedasticity (GARCH) or Stochastic Volatility (SV) models. Simulation experiments in Fisher and Gallagher (2012) show the weighting structure improves over the other methods available in the literature. The versatility of the method is demonstrated further as it can be used to detect nonlinear processes (i.e. McLeod and Li (1983)) and for a fitted GARCH process (i.e. Li and Mak (1994)).

Current and Future Work

My current and future work revolves around extending the time series goodness-of-fit test for other areas of time series and different testing procedures.

General Weighting Schemes

The results of Fisher and Gallagher (2012) suggest that decreasing weighted combinations of the autocorrelation function may be less sensitive to the choice of m. In practice, the lag at which the statistics are calculated will generally increase with the sample size. This motivates collaborator Colin Gallagher and I to consider the asymptotic behavior of general weighted portmanteau statistics:

$$Q_{n,m} = n \sum_{k=1}^{m} w_k \hat{r}_k^2, \tag{5}$$

where $\{w_k\} = \{w_{k,m}\}$ is a sequence of positive real numbers. Three results can be proven: (1) For fixed m, as $n \to \infty$, the general weighted Portmanteau statistic has an asymptotic distribution that can be represented as a weighted sum of independent chi square random variables each with one degree of freedom; (2) If m depends on n and the sum of the weights is divergent, then the test statistic can be normalized to be $O_p(1)$, the normalized test statistic has an asymptotic normal distribution; (3) if m diverges with the sample size and the weights are summable, then the statistic does not require normalization and the asymptotic distribution has weak limit which results from taking the limit as $m \to \infty$ of the fixed m asymptotic distribution.

We introduce two potential weighting schemes satisfying Theorem (3) and study their behavior through simulation. The correlation function of an ARMA process is summable in the lag and can be bounded by a constant multiple of a sequence which decreases geometrically in the lag. In other words the large lag correlations decay very quickly to zero. It seems intuitive that the weights in (5) be selected to decay quickly as well, since under the alternative hypothesis of an under fit model the correlations far from lag zero should still be relatively small. Based on this line of thought, we consider weights of the form $w_k = (p+q)a^{k-1}$, for some $0 < a < 1$. The constant multiplier $p + q$ is included to ensure that the second moment approximation is positive. The geometric decaying weights effectively truncate the statistic at some lag. For instance, with a ratio of $a = 0.9$ and $p + q = 1$ the weight at lag 31 is approximately 0.04. All squared autocorrelations beyond lag 31 are essentially not included in the statistic. Our simulation studies show this appears to stabilize the statistic as its empirical size and power appear constant as the lag increases. The geometric decaying scheme does pose a problem with no clear solution: How does one select the ratio a? The following adaptive weighting scheme eliminates the need to pick a ratio and appears to be more powerful.

The data-driven weights can be better understood through an example: let $X_t = \epsilon_t + \theta_5 \epsilon_{t-5}$. The data follows a seasonal moving-average of order 5 where the only correlation is at lag 5. The autocorrelation functions should deviate from zero only at lags $5, 10, \ldots$ In theory, we should put the most weight on those particular squared autocorrelations. Using one of the open questions in the discussion in Fisher and Gallagher (2012) we propose a combination of the autocorrelation function and the partial autocorrelations. Essentially use a function of the partial autocorrelations as the weights. Consider taking a linear combination of the Ljung Box statistic and a weighted statistic:

$$Q_W = n \sum_{k=1}^{m_0} \frac{n+2}{n-k} \hat{r}^2 \quad + \quad n \sum_{k=m_0+1}^{m} w_k \hat{r}^2.$$

The first m_0 terms get the standardizing weight $(n + 2)/(n - k)$ and the remaining $m - m_0$ terms will get absolutely summable weights. We suggest the later terms be given weights picked as a function of the partial autocorrelation that converge to zero in probability. Utilizing this scheme, for any bounded m_0, the weights will be absolutely summable in probability but will extract more information from the data at higher lags under the alternative hypothesis. Based on the work of Keenan (1997), we suggest $m_0 = \min(log(n), M)$ where M is a bounded constant and $w_k = -log(1 - |\pi_k|)$.

Initial simulation results suggest the proposed adaptive weighting scheme based on the data appears to dominate the competition. The proposed weighting schemes have almost constant empirical size and power as the lag increases for fixed n. These simulations have motived us to consider other cases and developing an adaptive techniques for detecting nonlinear processes. We expect to submit this result in Spring 2013.

Multivariate Time Series

A natural extension for a weighted statistic will be in the modeling of multivariate time series. The Vector ARMA (VARMA) process of a k-dimensional time series $\mathbf{X}_t = (X_{1t}, X_{2t}, \ldots, X_{kt})'$ is the natural analog of the univariate ARMA model into k-dimensions. After fitting such a process, one much check for the adequacy of the fitted model. The portmanteau statistics of Box and Pierce (1970) and Ljung and Box (1978) were extended to the multivariate setting by Chitturi (1974), Hosking (1980, 1981b) and Li and McLeod (1981). Hosking (1981a) noted that these statistics are equivalent. Each of these statistics are asymptotically distributed as chi square random variables with $k^2(m-p-q)$ degrees of freedom when p autoregressive and q moving average matrix parameters are fit to the k-dimensional time series and tested at lag m. Recently, Mahdi and McLeod (2012) generalized the statistics from Peña and Rodríguez (2002, 2006) for fitted VARMA.

Along with collaborator Michael Robbins, I am currently working on this topic and have preliminary results extending the methodology from Fisher and Gallagher (2012) for multivariate time series. Like that in Mahdi and McLeod (2012), considering the Toeplitz matrix on the multivariate autocorrelation matrix where each *element* is a matrix of the form $\widehat{\mathbf{R}}_k$ defined in Hosking (1980). Using the methodology from Fisher and Gallagher (2012) and properties of the trace operation, we have derived a statistic of the form

$$\mathbf{Q}_W = n^2 \sum_{k=1}^{m} \frac{m-i+1}{m} \text{tr}\left(\widehat{\mathbf{R}}_k' \widehat{\mathbf{R}}_k\right) / (n-k), \tag{6}$$

where m is the number of lags being tested. This statistic is asymptotically distributed as a linear combination of $k^2 m$ chi square random variables, each with one degree of freedom and the coefficients are the eigenvalues of the covariance matrix of $\mathbf{Q}_W$. We are currently exploring data-driven weighting schemes and looking to apply this methodology to other multivariate time series models (detecting and fitting GARCH).

Infinite Variance Time Series

Due to empirical findings the importance of infinite variance models has increased significantly. It may be the case that the innovation sequence comes from a heavy-tailed, extreme value, or infinite variance distribution. Runde (1997) provided the distribution of the popular Box Pierce statistic for time series with infinite variance based on those results. Lin and McLeod (2008) and Lee and Ng (2010) provide reviews of the topic and propose some portmanteau statistics. The test in Lin and McLeod (2008) is the logical adaption of the Peña and Rodríguez (2002) statistic into infinite variance time series. Lee and Ng (2010) suggest a modification of the Ljung-Box statistic where the residuals are trimmed at a suitable threshold. Bouhaddioui and Ghoudi (2012) develops a nonparametric test based on the rank of the correlations. Simulations show the newly suggested methods improve over the traditionally used Ljung Box type test.

The results of Davis and Resnick (1986), Lin and McLeod (2008), Lee and Ng (2010) and Bouhaddioui and Ghoudi (2012) can all be modified to utilize a weighting scheme as described above. I am currently working with collaborators Yunwei Cui and Rongning Wu on diagnostic procedures for noncausal time series (observation may depend on future innovation terms) with infinite variance, see Andrews et al. (2009) for model description. The ideas will be extended to study nonlinear models with infinite variance; an area relatively in its infancy as work began about 15-years ago. We are currently developing adaptive methodology based on the above results.

References

Andrews, B., Calder, M., Davis, R. A., 2009. Maximum likelihood estimation for α-stable autoregressive processes. Ann. Statist. 37 (4), 1946–1982.
URL http://dx.doi.org/10.1214/08-AOS632

Bouhaddioui, C., Ghoudi, K., 2012. Empirical processes for infinite variance autoregressive models. Journal of Multivariate Analysis 107 (0), 319 – 335.
URL http://dx.doi.org/10.1016/j.jmva.2012.01.018

Box, G. E. P., Pierce, D. A., 1970. Distribution of residual autocorrelations in autoregressive-integrated moving average time series models. J. Amer. Statist. Assoc. 65, 1509–1526.

Chitturi, R. V., 1974. Distribution of residual autocorrelations in multiple autoregressive schemes. J. Amer. Statist. Assoc. 69, 928–934.

Davis, R., Resnick, S., 1986. Limit theory for the sample covariance and correlation functions of moving averages. Ann. Statist. 14 (2), 533–558.
URL http://dx.doi.org/10.1214/aos/1176349937

Fisher, T. J., 2012. On testing for an identity covariance matrix when the dimensionality equals or exceeds the sample size. Journal of Statistical Planning and Inference 142 (1), 312 – 326.
URL http://dx.doi.org/10.1016/j.jspi.2011.07.019

Fisher, T. J., Gallagher, C. M., 2012. New weighted portmanteau statistics for time series goodness of fit testing. Journal of the American Statistical Association 107 (498), 777–787.
URL http://amstat.tandfonline.com/doi/abs/10.1080/01621459.2012.688465

Fisher, T. J., Sun, X., 2011. Improved Stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. Comput. Statist. Data Anal. 55 (5), 1909–1918.
URL http://dx.doi.org/10.1016/j.csda.2010.12.006

Fisher, T. J., Sun, X., Gallagher, C. M., 2010. A new test for sphericity of the covariance matrix for high dimensional data. J. Multivariate Anal. 101 (10), 2554–2570.
URL http://dx.doi.org/10.1016/j.jmva.2010.07.004

Hosking, J. R. M., 1980. The multivariate portmanteau statistic. J. Amer. Statist. Assoc. 75 (371), 602–608.

Hosking, J. R. M., 1981a. Equivalent forms of the multivariate portmanteau statistic. J. Roy. Statist. Soc. Ser. B 43 (2), 261–262.

Hosking, J. R. M., 1981b. Lagrange-multiplier tests of multivariate time-series models. J. Roy. Statist. Soc. Ser. B 43 (2), 219–230.

Keenan, D. M., 1997. A central limit theorem for $m(n)$ autocovariances. J. Time Ser. Anal. 18 (1), 61–78.
URL http://dx.doi.org/10.1111/1467-9892.00039

Ledoit, O., Wolf, M., December 2003. Improved estimation of the covariance matrix of stock returns with an application to portfolio selection. Journal of Empirical Finance 10 (5), 603–621.
URL http://dx.doi.org/10.1016/S0927-5398(03)00007-0

Lee, S., Ng, C. T., 2010. Trimmed portmanteau test for linear processes with infinite variance. J. Multivariate Anal. 101 (4), 984–998.
URL http://dx.doi.org/10.1016/j.jmva.2009.09.016

Li, W. K., Mak, T. K., 1994. On the squared residual autocorrelations in non-linear time series with conditional heteroskedasticity. J. Time Ser. Anal. 15 (6), 627–636.
URL http://dx.doi.org/10.1111/j.1467-9892.1994.tb00217.x

Li, W. K., McLeod, A. I., 1981. Distribution of the residual autocorrelations in multivariate ARMA time series models. J. Roy. Statist. Soc. Ser. B 43 (2), 231–239.

Lin, J.-W., McLeod, A. I., 2008. Portmanteau tests for ARMA models with infinite variance. J. Time Ser. Anal. 29 (3), 600–617.
URL http://dx.doi.org/10.1111/j.1467-9892.2007.00572.x

Ljung, G. M., Box, G. E. P., 1978. On a measure of lack of fit in time series models. Biometrika 65 (2), 297–303.

Mahdi, E., McLeod, I. A., 2012. Improved multivariate portmanteau test. Journal of Time Series Analysis 33 (2), 211–222.
URL http://dx.doi.org/10.1111/j.1467-9892.2011.00752.x

McLeod, A. I., Li, W. K., 1983. Diagnostic checking ARMA time series models using squared-residual autocorrelations. J. Time Ser. Anal. 4 (4), 269–273.
URL http://dx.doi.org/10.1111/j.1467-9892.1983.tb00373.x

Monti, A. C., 1994. A proposal for a residual autocorrelation test in linear models. Biometrika 81 (4), 776–780.
URL http://dx.doi.org/10.1093/biomet/81.4.776

Peña, D., Rodríguez, J., 2002. A powerful portmanteau test of lack of fit for time series. J. Amer. Statist. Assoc. 97 (458), 601–610.
URL http://dx.doi.org/10.1198/016214502760047122

Peña, D., Rodríguez, J., 2006. The log of the determinant of the autocorrelation matrix for testing goodness of fit in time series. J. Statist. Plann. Inference 136 (8), 2706–2718.
URL http://dx.doi.org/10.1016/j.jspi.2004.10.026

Runde, R., 1997. The asymptotic null distribution of the Box-Pierce Q-statistic for random variables with infinite variance. An application to German stock returns. J. Econometrics 78 (2), 205–216.
URL http://dx.doi.org/10.1016/S0304-4076(96)00008-5

Srivastava, M. S., 2005. Some tests concerning the covariance matrix in high dimensional data. J. Japan Statist. Soc. 35 (2), 251–272.

Stein, C., 1956. Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In: Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, 1954–1955, vol. I. University of California Press, Berkeley and Los Angeles, pp. 197–206.