1-1-2011

Non linear estimation of returns on hedge funds with scarce observations

Akram Samarikhalaj
Ryerson University

Non Linear Estimation of Returns on Hedge Funds with Scarce Observations

by

AKRAM SAMARIKHALAJ

Bachelor of Science, Shahrood University Of Thechnology, 1999

A thesis
presented to Ryerson University
in partial fulfillment of the
requirements for the degree of
Master of Science
in the program of
Applied Mathematics

Toronto, Ontario, Canada, 2011

Author's Declaration

I hereby declare that I am the sole author of this thesis.
I authorize Ryerson University to lend this thesis to other institutions or individuals for the purpose of scholarly research.

I further authorize Ryerson University to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research.

Abstract

Non Linear Estimation of Returns on Hedge Funds with Scarce Observations

Master of Science 2011
Akram Samarikhalaj
Applied Mathematics
Ryerson University

Explaining the behavior of a financial portfolio like a Hedge Fund is challenging for many reasons, one of those reasons is scarce observations. One possibility to circumvent these issues is to find simple relationships between the portfolio and financial factors. These factors are observed more frequently so it is valid to assume that one can estimate not only the conditional expectation with respect to single factors, but also the joint law of all the underlying factors. The problem, then, is to recover the conditional expectation of the portfolio's return given *all* the factors.The author of the paper,"Measuring Risk With Scarce Observation" prescribes a reasonable criteria which provides existence and uniqueness to this problem also characterizes the solution under the assumption of Gaussian distribution among the factors(Independent factors). In our thesis, we present a solution for the case when the joint law of factors is a multivariate t-student distribution.

Acknowledgments

I would like to thank my supervisor, Dr. Marcos Escobar, and my co supervisor, Dr. Sebastian Ferrando, who guided me along the way. It would not have been possible to complete this project without their help and advice.

I would also like to thank my parents and my sisters for their support during my studies at Ryerson University.

Contents

1 Introduction:

Our interest is a special type of investment companies called hedge funds. These are loosely regulated companies which could invest in a variety of complicated products, making their performance different to that of common stocks. These companies report returns (Y) on a monthly basis leading to scarce data bases therefore making statistical analysis more challenging. In general, investments are interested in many financial objectives related to hedge funds. For example, they may want to measure the risk of a Hedge fund and therefore trying to explain the returns (Y) and the variation in Y in terms of the variations of a set of variables X which could represent macroeconomic variables, fundamentals of a company or simple stocks and indexes. In general these variables are called factors for simplicity. Note that to measure the risk of a financial product with scarce observation, the simplest way is to relate it to the values of certain financial factors which are more popular or stable therefore leading to a more robust analysis. They also may want to find ways to hedge the performance of hedge funds companies on which they may have large investment allocations. This hedging exercise protect them, in particular, against downward movements of the returns of the hedge funds companies in their portfolios. A hedging exercise could be created by investing on common stocks, indexes (X) or even on financial derivatives ($f(X)$), the later are basically nonlinear functions on the underlying stocks and therefore can be seen as quadratic or higher order functions. This means that these investment companies would like to know how to combine allocations on polynomial functions of separate stocks ($g\left(f_1(X_1), ..., f_N(X_N)\right)$) in order to hedge the hedge fund returns (Y).

The mathematical problem would be easily solvable via regression analysis assuming enough data is available not only for the factors (X) to be used in the hedging but also for the hedge funds to be hedged. This later conditions is the one that fails as there are only dozens of data points available from hedge funds companies which come from based monthly performances (Y). On the other hand stocks or indexes are pretty much available on a daily or even intra-day basis making the analysis of their relationships ($X = (X_1, ..., X_N)$) easier to describe as opposed to describing the relationship between the hedge funds and the stocks ($Y = g(X)$).

In this context a recent paper: "Measuring Risk With Scarce Observation" [1], proposed an alternative approach to standard regression analysis with the purpose

of building an optimal multidimensional function g. The method proposed by the authors uses strongly the joint probability distribution of the factors X and put forward a new concept of optimality based on a two steps approach: it first finds the best relationship between each marginal factor (X_i) and the dependent variable Y, this is achieved via standard regression analysis therefore minimizing the fitting error. In a second step it looks for the multivariate function g with minimum variance such that the marginal fittings are satisfied. The requirement of minimum variance, is a way of finding a reasonably well behaved multidimensional function among all possible candidates avoiding at the same time the use of the joint distribution of X and the scarce variable Y.

This thesis studies in detail the paper by [1] and the statistical relationships and methods provided in there. It covers the topic of regression (linear and non-linear), which is one of the steps in the targeted methodology. In this case the regression could be performed between some financial factors and the returns of hedge fund companies. The statistical relationship between the factors altogether and a set of given hedge fund returns will be studied using the second step in [1]. Even though the authors provide some theoretical results about this two steps procedure, they fail to completely provide a methodology that could be used beyond the two simple cases they managed in their examples. The cases they used as examples were those when the factors X follow a multivariate Gaussian distribution (also a multivariate Gaussian Copula or dependence structure was studied) and the case where the factors were assumed independent.

In finance and economics Gaussian assumptions among variables are extremely unrealistic. This is due to the presence of asymmetries (skewness), high probability of extreme events (fat tails) and the presence of tail dependence as a non Gaussian copula feature. One of the most popular non-Gaussian random variable is the t-student, this is because the distribution satisfies some of the previously mentioned stylized facts on its univariate and multivariate variants. In order to adapt the aforementioned paper [1] to a context beyond Gaussian, several changes were performed. Among them, finding polynomials that has orthogonality relationship under a given measure can be a first step. This leads to some drawbacks in terms of the family of functions in which the optimal solution is found as well as on the possible marginal regression fittings that were compatible with the methodology. This modified analysis was then applied to the case of a univariate and therefore a multivariate extension of the t-student distribution. Some other possible families

of probability measures and the associated orthogonal polynomials were also mentioned.

The thesis is organized as follows: the next section provides an overview of hedge funds and regression techniques (linear and nonlinear). Chapter 2 reviews the most important results from [1] with emphasis on the results that will be extended or modified. Chapter 3 provides the novel results in the thesis, starting with Section 3.1 and the modifications to the existing framework as well as the methodology to build solutions. In Section 3.2, the application of the results from the previous section are developed. This involves defining the orthogonal polynomials under a t-student distribution, then selecting an appropriate multivariate t-student distribution and considering inner products under this measure. After that, we study in detail three particular cases which correspond to the smallest degrees of freedom, and therefore, representing the cases farther away from the Gaussian measure. We also consider the general case of any given number of degrees of freedom. Section 3.3 motivates some other measures and orthogonal functions for future research. Chapter 4 concludes.

1.1 Hedge Funds

A hedge fund is a fund that can take both long and short positions, look for arbitrage opportunities, buy and sell undervalued securities, trade options or bonds, and invest in almost any opportunity in any market where it foresees impressive gain at reduced risk. The primary aim of most hedge funds is to reduce risk while attempting to preserve capital and deliver positive returns under all market conditions.
Hedge funds are investment vehicles that explicitly pursue absolute returns on their underlying investments. The description "Absolute Return Fund" would be more accurate, since not all hedge funds contain an explicit hedge on their portfolio of investments. However the "Hedge Fund" definition has come to incorporate, any absolute return fund investing within the financial markets (stocks, bonds, commodities, currencies, derivatives, etc) and/or applying non-traditional portfolio management techniques including, but not restricted to, shorting, leveraging, arbitrage, swaps, etc. Hedge funds can invest in any number of strategies. These are perhaps identifiable by their structure, a limited partnership (the manager acting as the general partner and investors acting as the limited partners) with

performance related fees, high minimum investment requirements and restrictions on types of investors and entry and exit periods [6].

Investors decide to allocate funds to hedge funds for several reasons:

1. To increase the return on the portfolio.

   Many hedge funds have performed well in both absolute and relative return to aggregate stock and bond returns which is enough to make them appealing to many investors.

2. To diversify the returns of assets within the portfolio.

   Diversification involves a statistic called correlation. Correlation is a single number that describes the degree of relationship between two or more variables. For example, A correlation of one means that the two numbers related and if one grows so does the other. Two assets in the same industry provide less risk reduction from diversification than a combination of unrelated companies. A well diversified portfolio combines the returns of many assets often with some effort devoted to identifying returns that are not correlated.

3. To reduce risk.

   Many hedge fund have lower risk than traditional assets [15].

Note that the return of the hedge fund is published monthly and it reduces the size of the sample to estimate. For example if there are two years history of a hedge fund then there are two dozen observations. The confidence (or prediction) interval is an estimate of an interval in which future observations will fall and often used in regression analysis. It is used to indicate the reliability of an estimate. A major factor determining the length of the confidence interval is the size of the sample used in the estimation procedure. For a smaller confidence interval more precise results will be obtained. The greater the sample size the smaller the size of confidence interval and the greater the number of variable the greater the size of confidence interval. In hedge fund analysis the size of the sample is small as the return is monthly and we have small data.

1.2 Regression Analysis

In statistics, regression analysis includes any technique used for modeling and analyzing several variables when the focus is on the relationship between a dependent variable and one or more independent variables. The relationship is expressed in the form of an equation or a model connecting the response or dependent variable and one or more explanatory or predictor variables. Regression analysis estimates the conditional expectation of the dependent variable given the independent variables. The conditional expectation is the average value of the dependent variable when the independent variables are held fixed. The estimation target is a function of the independent variables called the regression function[16].

A large body of techniques for carrying out regression analysis has been developed. Familiar methods such as linear regression and ordinary least squares regression are parametric. This means that a regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.

The performance of regression analysis methods in practice depends on the form of the data-generating process, and how it relates to the regression approach being used. Since the true form of the data-generating process is not known, regression analysis depends to some extent on making assumptions about this process. These assumptions are sometimes (but not always) testable if a large amount of data is available. Regression models for prediction are often useful even when the assumptions are moderately violated, even though they may not perform optimally. In many applications, especially with small effects or questions of causality based on observational data, regression methods give misleading results[20].

Regression models involve the following terms:

1 - The unknown parameters, β. It can be a scalar or a vector.

2 - The independent variables, X.

3 - The dependent variable, Y.

A regression model relates Y to a function of X and β.

$$Y \approx f(X, \beta)$$

The approximation is usually formalized as $E(Y|X) = f(X, \beta)$. To carry out parametric regression analysis, the form of the function f must be specified. Sometimes the form of this function is based on knowledge about the relationship between Y and X that does not rely on the data. If no such knowledge is available, a flexible or convenient form for f is chosen.

Assume now that the vector of unknown parameters β is of length k. In order to perform a regression analysis the user must provide information about the dependent variable Y:

- If n data points of the form (Y, X) are observed, where $n < k$, most classical approaches to regression analysis cannot be performed, since the system of equations defining the regression model is undetermined, there is not enough data to recover β.

- If exactly $n = k$ data points are observed, and the function f is linear, the equations $Y = f(X, \beta)$ can be solved exactly rather than approximately. This reduces to solving a set of N equations with N unknowns (the elements of β) , which has a unique solution as long as the X are linearly independent. If f is nonlinear, a solution may not exist, or many solutions may exist.

- The most common situation is where $n > k$ data points are observed. In this case, there is enough information in the data to estimate a unique value for β that best fits the data in some sense.

In the last case, the regression analysis provides the tools for:

Finding a solution for unknown parameters β that will, for example, minimize the distance between the measured and predicted values of the dependent variable Y (also known as method of least squares).

Under certain statistical assumptions, regression analysis uses the surplus of information to provide statistical information about the unknown parameters β and predicted values of the dependent variable Y [20].

Decomposition property: Any random variable y can be expressed as

$$y = E(y|x) + \varepsilon$$

where $E(y|x)$ is the expected value of y for given values of random variable X and ε is a random variable satisfying

i) $E(\varepsilon|x) = 0$

ii)$E(\varepsilon h(x)) = 0$ where $h(.)$ is any function of x.

This means any variable can be decomposed in two parts: conditional expectation and orthogonal error term.

Prediction property:

Let m(x) be any function of x. Then

$$E(y|x) = argmin_{m(x)} E[(y - m(x))^2]$$

intuition : the conditional expectation is the best prediction where ' best ' means minimum mean squared error.

Proof 1.1 $(y - m(x))^2 = [(y - E(y|x)) - (E(y|x) - m(x))]^2 = (y - E(y|x))^2 + (E(y|x) - m(x))^2 - 2(y - E(y|x))(E(y|x) - m(x))$

- *The first term is not affected by the choice of m(x).*
- *The third term* $(y - E(y|x))(E(y|x) - m(x)) = \varepsilon(x)h(x)$ *and* $E(\varepsilon(x)h(x)) = 0$ *by the decomposition property.*

Hence the whole expression is minimized if $m(x) = E(y|x)$ *see [18].*

1.2.1 Linear Regression:

In linear regression, data is modeled using linear functions, and unknown model parameters are estimated from the data.

Given a data set $\{y_i, x_{i1}, \ldots, x_{ik}\}_{i=1}^n$ of n statistical units, a linear regression model assumes that the relationship between the dependent variable y_i and the k-vector of independent variables x_i is linear.

This relationship is modeled through a term ε_i, an unobserved random variable that adds noise to the linear relationship between the dependent variable and independent variables. Thus the model takes form:

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ik} + \varepsilon_i = x_i'\beta + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $x_i^{'}$ is the ith. (row) vector of predictors for n observation and β is the vector of regression parameters to be estimated and ε_i is a random error. $'$ denotes the transpose, so that $x_i'\beta$ is the inner product between vectors x_i and β.

Some remarks on general use:

- y_i, is called dependent variable.
- The decision as to which a variable in a data set is modeled as the dependent variable and which are modeled as the independent variables may be based on a presumption that the value of one of the variables is caused by, or directly influenced by the other variables. Alternatively, there may be an operational reason to model one of the variables in terms of the others, in which case there need be no presumption of causality.
- x_i, are called predictor variables, or independent variables.
- Usually a constant is included as one of the independent variables. For example we can take $x_{i1} = 1$ for $i = 1, ..., n$. The corresponding element of β is called the intercept. Many statistical inference procedures for linear models require an intercept to be present, so it is often included even if theoretical considerations suggest that its value should be zero.

- Sometimes one of the independent variables can be a non-linear function of another independent variables or of the data, as in polynomial regression and segmented regression. The model remains linear as long as it is linear in the parameter vector β.
- The independent variables x_i may be viewed either as random variables, which we simply observe, or they can be considered as predetermined fixed values which we can choose.
- β, is a k-dimensional parameter vector. Its elements are also called regression coefficients. ε_i, is called the error term, or noise. This variable captures all other factors which influence the dependent variable y other than the independent variables x_i.The relationship between the error term and the independent variables, for example whether they are correlated, is a crucial step in formulating a linear regression model, as it will determine the method to use for estimation [13].

Assumptions of linear regression:

There are some principal assumptions which justify the use of linear regression models for purposes of prediction:

(1) Linearity of the relationship between dependent and independent variables (linearity on the parameters).

Violations of linearity are extremely serious if we fit a linear model to data which are nonlinearly related, our predictions are likely to be seriously in error, especially when we extrapolate beyond the range of the sample data.

(2) Independence of the errors (no serial correlation).

The errors also assumed to be uncorrelated across observations, so that for two observations i and j, the covariance between ε_i and ε_j is zero.
Violations of independence are also very serious in time series regression models: serial correlation in the residuals means that there is room for improvement in the model, and extreme serial correlation is often a symptom of a badly miss-specified model. Serial correlation is also sometimes a by-product of a violation of the linearity assumption–as in the case of a simple (i.e., straight) trend line fitted to data

which are growing exponentially over time.

(3) Homoscedasticity (constant variance) of the errors.

The errors are assumed to be homoscedastic, which means that for a given x, the errors have a constant variance. Formally ,

$$Var(\varepsilon_i|x_i) = \sigma^2 \quad for\ all\ i.$$

When the variance differs across observations, the errors are heteroscedastic and $Var(\varepsilon_i|x_i) = \sigma_i^2 \quad for\ all\ i$

Violations of homoscedasticity make it difficult to gauge the true standard deviation of the forecast errors, usually resulting in confidence intervals that are too wide or too narrow. In particular, if the variance of the errors is increasing over time, confidence intervals for out-of-sample predictions will tend to be unrealistically narrow. Heteroscedasticity may also have the effect of giving too much weight to small subset of the data (namely the subset where the error variance was largest) when estimating coefficients.

(4) Normality of the error distribution.

Violations of normality compromise the estimation of coefficients and the calculation of confidence intervals. Sometimes the error distribution is "skewed" by the presence of a few large outliers. Since parameter estimation is based on the minimization of squared error, a few extreme observations can exert a disproportionate influence on parameter estimates. Calculation of confidence intervals and various significance tests for coefficients are all based on the assumptions of normally distributed errors. If the error distribution is significantly non-normal, confidence intervals may be too wide or too narrow.

(5) Zero Condition Mean of ε.

$$E(\varepsilon_i|x_i) = 0.$$

(6) The x's are linearly independent.
This means that none of the x's is a linear combination of remaining x's.
If any of these assumptions is violated (i.e., if there is nonlinearity, serial correlation, heteroscedasticity, and/or non-normality), then the forecasts, confidence

intervals, and economic insights yielded by a regression model may be (at best) inefficient or (at worst) seriously biased or misleading [13].

1.2.2 Non-Linear Regression:

The basic idea of nonlinear regression is the same as that of linear regression. Nonlinear regression is characterized by the fact that the prediction equation depends nonlinearly on one or more unknown parameters. Whereas linear regression is often used for building a purely empirical model, nonlinear regression usually arises when there are physical reasons for believing that the relationship between the response and the predictors follows a particular functional form.
In the more general normal nonlinear regression model, the function $f(.)$ relating the response to the predictors is not necessarily linear:

$$y_i = f(x_i, \beta) + r_i$$

As in linear model, β is a vector of parameters and x_i is a vector of predictors (but in the nonlinear regression model, these are not generally of the same dimension)and the r_i are random errors.

In nonlinear regression the data is fitted by a method of successive approximation [13].

Assumptions of nonlinear regression:
There are some principal assumptions which justify the use of non-linear regression models for purposes of prediction:

(1) The model is correct. Nonlinear regression adjusts the variables in the equation you chose to minimize the sum-of-squares. It does not attempt to find a better equation.
(2) The variability of values around the curve follow a Gaussian distribution. Even though no biological variable follows a Gaussian distribution exactly, it is sufficient that the variation be approximately Gaussian.
(3) Homoscedasticity (constant variance) of the errors.
The errors are assumed to be homoscedastic, which means that for a given x, the errors have a constant variance. Formally ,

$$Var(\varepsilon_i|x_i) = \sigma^2 \quad for\ all\ i.$$

It means the SD (standard deviation) of the variability is the same everywhere, regardless of the value of X. The assumption is termed homoscedasticity. If the SD is not constant but rather is proportional to the value of Y, you should weight the data to minimize the sum-of-squares of the relative distances.
(4) The model assumes that you know X exactly. This is rarely the case, but it is sufficient to assume that any imprecision in measuring X is very small compared to the variability in Y.
(5) The errors are independent. The deviation of each value from the curve should be random, and should not be correlated with the deviation of the previous or next point. If there is any carryover from one sample to the next, this assumption will be violated [13].

1.2.3 Least Squares

The method of least squares is a standard approach to the approximate solution of overdetermined systems, i.e. sets of equations in which there are more equations than unknowns. "Least squares" means that the overall solution minimizes the sum of the squares of the errors made in solving every single equation.
The most important application is in data fitting. The best fit in the least-squares sense minimizes the sum of squared residuals, a residual being the difference between an observed value and the fitted value provided by a model. Least squares problems fall into two categories: linear or ordinary least squares and non-linear least squares, depending on whether or not the residuals are linear in all unknowns. The linear least-squares problem occurs in statistical regression analysis. The non-linear problem has no closed-form solution and is usually solved by iterative refinement; at each iteration the system is approximated by a linear one, thus the core calculation is similar in both cases.
The method of least squares assumes that the best-fit curve of a given type is the curve that has the minimal sum of the deviations squared (least square error) from a given set of data.
Suppose that the data points are $(x_1, y_1), ..., (x_n, y_n)$ where x is the independent variable and y is the dependent variable. The fitting curve $f(x)$ has the error d for each data point, i.e. $d_1 = y_1 - f(x_1), d_2 = y_2 - f(x_2), ..., d_n = y_n - f(x_n)$.

According to the method of least squares, the best fitting curve has the property that:

$$\Pi = d_1^2 + d_2^2 + ... + d_n^2 = \sum_{i=1}^{n} d_1^2 = \sum_{i=1}^{n} [y_i - f(x_i)]^2 \rightarrow min$$

Polynomials are one of the most commonly used curves in regression. When using an mth degree polynomial

$$y = a_0 + a_1 x + a_2 x^2 + ... + a_m x^m$$

to approximate the given set of data, $(x_1, y_1), ..., (x_n, y_n)$, where $n \geq m + 1$, the best fitting curve $f(x)$ has the least square error, i.e.,

$$\Pi = \sum_{i=1}^{n} [y_i - f(x_i)]^2 = \sum_{i=1}^{n} [y_i - (a_0 + a_1 x_i + a_2 x_i^2 + ... + a_m x_i^m)]^2 = min$$

Note that $a_0, a_1, ..., a_m$ are unknown coefficients while all x_i and y_i are given. The unknown coefficients can be obtained by solving linear equations below [11].

$$\partial \Pi / \partial a_j = 0 \quad j = 1, ..., m$$

Solving linear least square Problem:
The general problem: Consider a system

$$\sum_{j=1}^{n} X_{ij} \beta_j = y_i, \quad (i = 1, 2, ..., m)$$

of m linear equations in n unknown coefficients $\beta_1, \beta_2, ..., \beta_n$, with $m > n$. This can be written in matrix form as

$$X\beta = Y$$

The goal is to find the coefficients β which fit the equations "best" in the sense of solving the quadratic minimization problem

$$\hat{\beta} = argminS(\beta)$$

where the objective function S is given by

$S(\beta) = \sum_{i=1}^{m} |y_i - \sum_{j=1}^{n} X_{ij}\beta_j|^2 = ||y - X\beta||^2.$

where $||\ .\ ||$ is the standard L^2-norm in the n-dimensional Euclidean space R^n. A justification for choosing this criterion is given in properties below. This minimization problem has a unique solution, provided that the n columns of the matrix X are linearly independent, given by solving the normal equations

$$(X^{'}X)\hat{\beta} = X^{'}y.$$

Define the ith residual to be

$$r_i = y_i - \sum_{j=1}^{n} X_{ij}\beta_j.$$

Then $S(\beta)$ can be rewritten

$$S(\beta) = \sum_{i=1}^{m} r_i^2$$

S is minimized when its gradient vector is zero. The elements of the gradient vector are the partial derivatives of S with respect to the parameters:

$$\frac{\partial S}{\partial \beta_j} = 2\sum_{i=1}^{m} r_i \frac{\partial r_i}{\partial \beta_j} \quad (j = 1, 2, ..., n).$$

The derivatives are

$$\frac{\partial r_i}{\partial \beta_j} = -X_{ij}.$$

Substitution of the expressions for the residuals and the derivatives into the gradient equations gives

$$\frac{\partial S}{\partial \beta_j} = 2\sum_{i=1}^{m} \left(y_i - \sum_{k=1}^{n} X_{ik}\beta_k \right) (-X_{ij}) \quad (j = 1, 2, ..., n).$$

Thus if $\hat{\beta}$ minimizes S, we have

$$2\sum_{i=1}^{m}\left(y_i - \sum_{k=1}^{n} X_{ik}\hat{\beta}_k\right)(-X_{ij}) = 0 \quad (j = 1, 2, ..., n).$$

Upon rearrangement, we obtain the normal equations:

$$\sum_{i=1}^{m}\sum_{k=1}^{n} X_{ij}X_{ik}\hat{\beta}_k = \sum_{i=1}^{m} X_{ij}y_i \ (j = 1, 2, ..., n).$$

The normal equations are written in matrix notation as

$$(X^{'}X)\hat{\beta} = X^{'}y.$$

The solution of the normal equations yields the vector $\hat{\beta}$ of the optimal parameter values [12].

Properties of the least-squares estimators:
The gradient equations at the minimum can be written as

$$(y - X\hat{\beta})X = 0$$

The vector of residuals, $y - X\hat{\beta}$ is orthogonal to the column space of X, since the dot product $(y - X\hat{\beta}).X$ is equal to zero. This means that $y - X\hat{\beta}$ is the shortest of all possible vectors $y - X\hat{\beta}$, that is, the variance of the residuals is the minimum possible[19].
The Gauss Markov theorem states that in a linear regression model in which **the errors have expectation zero and are uncorrelated and have equal variances** , the best linear unbiased estimator of the coefficients is given by the ordinary least squares estimator. Here "best" means giving the lowest possible mean squared error of the estimate. The errors need not be normal, nor independent and identically distributed (only uncorrelated and homoscedastic).see [21]
Limitations and Alternatives:
The independent variable, x, is free of error.

In practice, the errors on the measurements of the independent variable are usually much smaller than the errors on the dependent variable and can therefore be ignored. When this is not the case, total least squares also known as errors-in-variables models, or rigorous least squares, should be used. This can be done by adjusting the weighting scheme to take into account errors on both the dependent and independent variables and then following the standard procedure.
In some cases the (weighted) normal equations matrix is ill-conditioned.
When fitting polynomials the normal equations matrix is a Vandermonde matrix. Vandermode matrices become increasingly ill-conditioned as the order of the matrix increases. In these cases, the least squares estimate amplifies the measurement noise and may be grossly inaccurate. Various regularization techniques can be applied in such cases, the most common of which is called ridge regression. If further information about the parameters is known, for example, a range of possible values of β, then various techniques can be used to increase the stability of the solution. Another drawback of the least squares estimator is the fact that the norm of the residuals, $||y - X\beta||$ is minimized, whereas in some cases one is truly interested in obtaining small error in the parameter β, e.g. , a small value of $||\beta - \hat{\beta}||$. However, since β is unknown, this quantity cannot be directly minimized. The least squares method is often applied when no prior is known. Surprisingly, when several parameters are being estimated jointly, better estimators can be constructed, an effect known as Stein's phenomenon.

Solving Nonlinear least squares Problem

Non-linear least squares is the form of least squares analysis which is used to fit a set of m observations with a model that is non-linear in n unknown parameters ($m_1 > n_1$). It is used in some forms of non-linear regression.

Consider a set of m data points, $(x_1, y_1), (x_2, y_2), ..., (x_{m_1}, y_{m_1})$, and a curve (model function) $y = f(x, \beta)$, that in addition to the variable x also depends on n parameters, $\beta = (\beta_1, \beta_2, ..., \beta_{n_1})$, with $m_1 \geq n_1$. It is desired to find the vectorβ of parameters such that the curve fits best the given data in the least squares sense, that is, the sum of squares:

$$S = \sum_{i=1}^{m_1} r_i^2$$

is minimized, where the residuals (errors) r_i are given by

$$r_i = y_i - f(x_i, \beta) \quad i = 1, 2, ..., m_1$$

The minimum value of S occurs when the gradient is zero. Since the model contains n parameters there are n gradient equations:

$$\frac{\partial S}{\partial \beta_j} = 2\sum_i r_i \frac{\partial r_i}{\partial \beta_j} = 0 \quad (j = 1, \ldots, n_1).$$

In a non-linear system, the derivatives $\frac{\partial r_i}{\partial \beta_j}$ are functions of both the independent variable and the parameters, so these gradient equations do not have a closed solution. Instead, initial values must be chosen for the parameters. Then, the parameters are refined iteratively, that is, the values are obtained by successive approximation,

$$\beta_j \approx \beta_j^{k+1} = \beta_j^k + \Delta\beta_j.$$

Here, k is an iteration number and the vector of increments, $\Delta\beta$, is known as the shift vector. At each iteration the model is linearized by approximation to a first-order Taylor series expansion about β^k
$f(x_i, \beta) \approx f(x_i, \beta^k) + \sum_j \frac{\partial f(x_i, \beta^k)}{\partial \beta_j} \left(\beta_j - \beta_j^k\right) \approx f(x_i, \beta^k) + \sum_j J_{ij}\Delta\beta_j.$
Where $J_{ij} = \frac{\partial f(x_i, \beta^k)}{\partial \beta_j}$ is the first-order partial derivatives of a function $f(x_i, \beta^k)$ with respect to β_j where $i = 1, 2, ..., m_1$ and $j = 1, \ldots, n_1$ and k is an iteration number,so J is a function of the independent variable and the parameters, so it changes from one iteration to the next. Thus, in terms of the linearized model, $\frac{\partial r_i}{\partial \beta_j} = -J_{ij}$and the residuals are given by:

$$r_i = \Delta y_i - \sum_{s=1}^{n_1} J_{is}\Delta\beta_s; \Delta y_i = y_i - f(x_i, \beta^k)$$

Substituting these expressions into the gradient equations, they become

$$-2\sum_{i=1}^{m_1} J_{ij}\left(\Delta y_i - \sum_{s=1}^{n_1} J_{is}\Delta\beta_s\right) = 0$$

which, on rearrangement, become n simultaneous linear equations, the normal equations

$$\sum_{i=1}^{m_1}\sum_{s=1}^{n_1} J_{ij}J_{is}\Delta\beta_s = \sum_{i=1}^{m_1} J_{ij}\Delta y_i \qquad (j = 1, \ldots, n_1).$$

The normal equations are written in matrix notation as (see [12]):

$$\left(\mathbf{J^T J}\right)\boldsymbol{\Delta}\beta = \mathbf{J^T}\boldsymbol{\Delta}\mathbf{y}.$$

Differences between LLSQ (linear least squares) and NLLSQ (non-linear least squares):
(1) The model function, f, in LLSQ (linear least squares) is a linear combination of parameters of the form $f = X_{i1}\beta_1 + X_{i2}\beta_2 + \cdots$ The model may represent a straight line, a parabola or any other linear combination of functions. In NLLSQ the parameters appear as functions, such as $\beta^2, e^{\beta}x$ and so forth. If the derivatives $\partial f/\partial\beta_j$ are either constant or depend only on the values of the independent variable, the model is linear in the parameters. Otherwise the model is non-linear.
(2) Algorithms for finding the solution to a NLLSQ problem require initial values for the parameters, LLSQ does not. Like LLSQ, solution algorithms for NLLSQ often require that the Jacobian be calculated. Analytical expressions for the partial derivatives can be complicated. If analytical expressions are impossible to obtain either the partial derivatives must be calculated by numerical approximation or an estimate must be made of the Jacobian.
(3) In NLLSQ non-convergence (failure of the algorithm to find a minimum) is a common phenomenon whereas the LLSQ is globally concave so non-convergence is not an issue.
(4) NLLSQ is usually an iterative process. The iterative process has to be terminated when a convergence criterion is satisfied. LLSQ solutions can be computed using direct methods.
(5) In LLSQ the solution is unique, but in NLLSQ there may be multiple minima in the sum of squares. Under the condition that the errors are uncorrelated with the predictor variables, LLSQ yields unbiased estimates, but even under that condition NLLSQ estimates are generally biased[21].

2 Cherny's Framework

2.1 Problem and General Solution

In Cherny's et al [1], the authors considered the problem of measuring the risk of a hedge fund with scarce observations. The simplest way of doing so is by taking the empirical distribution but with scarce data this could be challenging. A more advanced procedure, proposed by the authors, consist in relating the return of the hedge fund to the values of certain financial factors, like the price of the oil or macroeconomic factors. These factors are observed more frequently so it is assumed that one can estimate the joint law of all factors from the available data. Moreover, if the marginal conditional expectation of the hedge fund with respect to each of the factors separately are known, like in a regression, then the problem become that of recovering the conditional expectation of the hedge fund's return given all financial factors from the joint law and the marginal conditional expectation.
For a fixed time period $[0, T]$, let R denote the return of the hedge fund over this period and let $(X_1, X_2, ..., X_N)$ be the returns of the factors over this period. In general, one useful mathematical problem linking R to $(X_1, X_2, ..., X_N)$ would be to estimate the conditional distribution $P(R|X_1, X_2, ..., X_N)$.

If the joint law (probability) of $(X_1, X_2, .., X_N)$ is known, then the above problem is equivalent to estimating the joint law of $(R, X_1, X_2, .., X_N)$. If the hedge fund has a two year history, then there are only two dozen observation of R. This may be sufficient to estimate R and the joint distribution $Law(R, X_n)$ for $n = 1, ..., N$ but the data would be too scarce to try to estimate $Law(R, X_i, X_j)$ or the Law of more than three variables simultaneously. A simpler problem would be to recover $Law(R, X_1, X_2, .., X_N)$ from the knowledge of $Law(X_1, X_2, .., X_N)$ and $Law(R, X_n)$ $n = 1, 2, .., N$ but this is still a challenging problem.

A simpler mathematical problem is that of recovering the conditional expectation $E(R|X_1 = x_1, X_2 = x_2, ..., X_N = x_n)$ instead of the $Law(R|X_1 = x_1, X_2 = x_2, ..., X_N = x_N)$. For that, the authors use the $Law(X_1, X_2, ..., X_N)$ and the marginal conditional expectation $E(R|X_n = x)$ instead of the $Law(R, X_n)$.
Here, we present their framework in the most general way. Details are provided next:

Let us assume a measure P on $\mathbb{R}^{\mathbb{N}}$ (this is the law of $(X_1, X_2, ..., X_N)$) and the functions $\phi_n : \mathbb{R} \to \mathbb{R}$ $[\phi_{\ltimes}$ means $E(R|X_n = x)]$ are given. The problem is to find a function $\phi : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}$ $(\phi(x_1, x_2, .., x_N)$ which has the meaning of $E(R|X_1 = x_1, X_2 = x_2, .., X_N = x_n))$ such that:

$$E(\phi(X_1, X_2, ..., X_N)|X_n) = \phi_n(X_n) \quad n = 1, 2, .., N \tag{1}$$

Here X_n denotes the n-th coordinate projection of $\mathbb{R}^{\mathbb{N}}$ on $\mathbb{R}$. In order to obtain a unique solution Cherny [1] imposed additional conditions on the function ϕ. They proposed to look for the solution which is the most moderate one as measured by its variance.

$$\begin{cases} Minimize \;\; var \; \phi(X_1, X_2, .., X_N) \\ E(\phi(X_1, X_2, ..., X_N)|X_n) = \phi_n(X_n) \qquad n = 1, 2, .., N, \end{cases} \tag{2}$$

where var denotes the variance and the minimization is performed with respect to a family of integrable functions ϕ which is defined according to a given measure P on the factors.

2.2 General Setup

In this section, we will study problem for an arbitrary measure P. Let us set

$$\Phi = \{(\phi_1, \phi_2, ..., \phi_N) : \; E\phi_n^2(X_n) < \infty \; and \; E\phi_n(X_n) = 0 \; \forall n = 1, ..., n\}$$

Let $Pr_{\mathbb{E}}$ denote the orthogonal projection on a space $\mathbb{E}$ and $||.||$ the $L^2 - norm$. The following lemma sheds light on the structure of the solutions. In page 11 of [1], the solution of the case when the distribution between of the factors is Gaussian, the application of next Lemma completed the proof.

Lemma 2.1 *Let $(\phi_1, \phi_2, ..., \phi_N) \in \Phi$ and suppose that $\psi_n : \Re \longrightarrow \Re$ are measurable functions with $E\psi_n^2(X_n) < \infty$ such that the function*

$$\phi(x_1, ..., x_N) = \sum_{n=1}^{N} \psi_n(x_n) \tag{3}$$

satisfies (1); then it is the unique solution of (2).

The following proof is a more detailed version of the proof in the paper.

Proof 2.2 *Denote*

$$E_n = \{\xi \in L^2 : \xi \;\; is \;\; X_n - measurable, E\xi = 0\}$$

Note X_n is a random variable define in a probability space $(\Omega^*, \mathfrak{F}, \mathfrak{P})$, therefore the function ξ is Xn-measurable iff $\xi = g(Xn)$ for some g such that the preimage of each measurable set (on the Borel Algebra on $\mathfrak{R}$) is in $\mathfrak{F}$ (so g is a measurable function).
Let us assume $\tilde{\phi}$ satisfies (1) , this means that

$$E(\tilde{\phi} \;|X_n) = \phi_n(X_n) = E(\phi|X_n)$$

As $E(\phi|X_n)$ is the projection of ϕ onto E_n then for any $Y \in E_n$:

$$< Y, \phi - E(\phi|X_n) >= \int Y(\phi - E(\phi|X_n))dP = 0$$

To see this,

$$\begin{aligned} \int Y(\phi - E(\phi|X_n))dp &= E(Y\phi) - E(YE(\phi|X_n)) \\ &= E(Y\phi) - E(E(Y\phi|X_n)) \;\; (since Y \in E_n) \\ &= E(Y\phi) - E(Y\phi) \end{aligned}$$

Therefore

$$Pr_{En}\tilde{\phi} = \phi_n(X_n) = Pr_{En}\phi$$

This implies that

$$\begin{aligned} Pr_{En}\tilde{\phi} - Pr_{En}\phi &= Pr_{En}(\tilde{\phi} - \phi) = 0 \\ n &= 1, ..., N \end{aligned}$$

So $(\tilde{\phi} - \phi)$ is orthogonal to E_n for $n = 1, ..., N$

$$< \tilde{\phi} - \phi, \xi_n >= 0, \forall, \xi_n \in E_n$$

as $\phi \in E_1 + E_2 + ... + E_N$ *(which is the space of sums* $\xi_1 + ... + \xi_N$ *where* $\xi_i \in E_i$ *) so* $\phi = \xi_1 + ... + \xi_N$ *we can imply* $\tilde{\phi} - \phi$ *is orthogonal to* ϕ*.*

To see this:

$$< \tilde{\phi} - \phi, \phi >=< \tilde{\phi} - \phi, \Sigma_{i=1}^{N} \xi_i >= \Sigma_{i=1}^{N} < \tilde{\phi} - \phi, \xi_i >= 0$$

This implies $||\tilde{\phi}|| \geq ||\phi||$*, which is a direct result from the following equations:*

$$\begin{aligned} \tilde{\phi} &= \phi - (\phi - \tilde{\phi}) \\ ||\tilde{\phi}||_2 &= ||\phi||_2 + ||(\phi - \tilde{\phi})||_2 \end{aligned}$$

Hence

$$||\tilde{\phi}||_2 \geq ||\phi||_2$$

and the equality is possible only if $\tilde{\phi} = \phi$*. This allows us to conclude that the linear combination minimizes the norm.*

The differences between non-linear regression and Cherny's et al solution is explained next. According to [1], R is the return of the hedge fund over a fixed period, $X_1, X_2, ...X_N$ are the returns of the factors over this period. The problem is to estimate the conditional expectation

$$E(R|X_1, ..., X_N)$$

from the marginal conditional expectations $E(R|X_n = x) \quad n = 1, ..., N$ and the joint probability for $(X_1, \ldots, X_N)$.

The authors assumed that $R, X_1, X_2, ...X_N$ are random variables with mean zero and $E(X_n)^2 = 1$, with R a dependent variable and $X_1, X_2, ...X_N$ the independent variables. In general a linear or nonlinear regression of R on each $X_n \quad , n = 1, ..., N$ would lead to:

$$R = \phi_i(X_i) + \varepsilon_i \quad i = 1, ..., N$$

where $\varepsilon_i's$ are errors. By the method of least square, the best-fit is the curve that has a minimal sum of the deviations squared (least square error) from X_n. So, the best $\phi_n(X_n) \quad n = 1, ..., N$ leads to a minimum value for

$$E(R - \phi_n(X_n))^2$$

As regression analysis estimates the conditional expectation of the dependent variable given the independent variables that is, the average value of the dependent variable when the independent variables are held fixed:

$$\phi_n(x) = E(R|X_n = x).$$

After finding the best fitted curves, [1] found a $E(R|X_1, ..., X_N)$ which is a function $\phi : \Re^N \to \Re$ and $\phi(X_1, ..., X_N) = E(R|X_1, ..., X_N)$ such that $E(\phi(X_1, ..., X_N)|X_n) = \phi_n(X_n)$.

The differences between [1] and a linear/nonlinear regression of R on $X_1, ..., X_N$ is that the former selects ϕ with the minimun variance among a wide family of possible choices. On the other hand regression minimizes the error $E(R - \phi(X_1, X_2, ..., X_N))^2$ based on a least square method for a specific set of parametric fucntions ϕ.

In other words, note that in a regression the error and the X variables are assumed independent (uncorrelated), hence from the expression:

$$R = \phi(X_1, ..., X_N) + \varepsilon$$

we could see that:

$$\begin{aligned} Var(R) &= Var(\phi(X_1, ..., X_N) + \varepsilon) \\ &= Var(\phi(X_1, ..., X_N)) + Var(\varepsilon) \end{aligned}$$

Therefore by minimizing the $Var(\phi(X_1, ..., X_N))$ (the target in [1]) the $Var(\varepsilon)$ will be maximized as variance of R is fixed, while in regression ϕ is chosen such that $Var(\varepsilon)$ is minimum.

$$Var(\varepsilon) = E([R - \phi(X_1, ..., X_N)]^2).$$

Concluding, the objectives are different between regression and [1], the former minimizes error within a narrow set of functions, while the later selects the solution among a wide set of possible fitting functions by requiring to have minimum variance.

2.3 Solution for Gaussian and Independent Measures

The solution of (2) for the case when the distribution between the financial factors is Gaussian has the form:

$$\phi(X_1, X_2, ..., X_N) = \sum_{n=1}^{N} \sum_{m=1}^{\infty} \alpha_{nm} H_m(X_n) \quad , x_n \in \Re$$

where α_{nm} are found through solving certain N-dimensional linear system and $H_m(X_n)$ are Hermite polynomials as Hermite polynomials have orthogonality relationship under gaussian measure.
One way to define Hermite polynomials is as follows:

$$f(a) = exp\{ax - \frac{a^2}{2}\}$$

$$H_m(x) = \frac{1}{\sqrt{m!}} \frac{\partial^m}{\partial a^m}|_{a=0} \; f(a) \quad x \in \Re$$

An an example,

$$H_0(x) = 1$$

$$H_1(x) = x$$

$$H_2(x) = \frac{(x^2 - 1)}{\sqrt{2}}$$

$$H_3(x) = \frac{(x^3 - 3x)}{\sqrt{6}}$$

Denote

$$a_{mn} = E[\phi_n(x_n) H_m(x_n)]$$

where the expectation is with respect to the Gaussian Measure with density function $f(x) = \frac{1}{\sqrt{2\pi}} \int_R e^{\frac{-x^2}{2}} dx$ so

$$\begin{aligned} a_{mn} &= \frac{1}{\sqrt{2\pi}} \int_R \phi_n(x) H_m(x) e^{\frac{-x^2}{2}} dx \\ n &= 1, 2, ..., N, m \in \mathbb{N} \end{aligned}$$

Denote by C the covariance matrix of $(X_1, X_2, ..., X_N)$ and C^m its m-th componentwise power. Each C^m is symmetric, positively definite, and non-degenerate. For each $m \in \aleph$, the vector

$$\begin{pmatrix} \alpha_{1m} \\ \vdots \\ \alpha_{Nm} \end{pmatrix} = C^{-1} \begin{pmatrix} a_{1m} \\ \vdots \\ a_{Nm} \end{pmatrix}$$

is well defined. (the proof is in [1] page 11). Suppose that $X_1, X_2, ..., X_N$ are independent under P then the solution of (2) is given by:

$$\phi(X_1, X_2, ..., X_N) = \sum_{n=1}^{N} \phi_n(X_n)$$

(the proof is in [1] page 10).

Independent components (special case of Gaussian). One special case of Gaussian is when $X_1, ..., X_N$ are Gaussian and independent. In this case $E(XY) = E(X)E(Y) = 0$ then $\rho = 0$ so according to lemma A.1 of [1] $\langle H_m(X), H_k(Y)\rangle = 0$ and also

$$\langle H_m(X), H_k(X)\rangle = \begin{cases} 1 & \text{if } m = k \\ 0 & \text{if } m \neq k \end{cases}$$

In page 11 of [1] we have

$$\phi_n(X_n) = \sum_{m=1}^{\infty}(\sum_{s=1}^{\infty}\sum_{k=1}^{N} \alpha_{ks}\langle H_s(X_k), H_m(X_n)\rangle)H_m(X_n)$$

$$= \sum_{m=1}^{\infty}(\alpha_{nm})H_m(X_n)$$

In the solution of the Gaussian case when we substitute $\phi_n(X_n) = \sum_{m=1}^{\infty}(\alpha_{nm})H_m(X_n)$ then we have

$$\phi(X_1, X_2, ..., X_N) = \sum_{n=1}^{N} \phi_n(X_n)$$

which is equal to the solution of (4.1) of [1].

3 Alternative Mathematical Framework

As seen in the previous section, where a summary of the results in [1] was provided, their method is applied to two simple joint measures, the independent measure and the multivariate Gaussian measure. The authors also explore the case of a Gaussian dependence structure, therefore a Gaussian Copula (see [1]). These examples are too simplistic for their targeted applications, which are financial variables. It is well known that financial instruments like stock prices, company's fundamentals and macroeconomic variables show non-gaussian features like fat tails (Kurtosis) in the marginals as well as tail dependence on the joint. One of the most popular distribution in the world of finance and economics, that allows for fat tails and tail dependence is the t-student and their multidimensional counterparts. This is why our main objective is to explore the applicability of the work developed by the previous authors beyond normality and in particular for a t-student case.

The mathematical setting used in [1] has several limitations which become clear once the method is applied to non gaussian measures like that of the t-student. For example, even though the authors show existence of the solution, they do not provide a methodology to build this solution given a probability measure, some hints can be extracted from their applications and this is one of the new developments in our thesis. Moreover, their main results strongly use the space of L^2 functions under the given probability measure, while the applications described [1] make use of a basis under this measure.

The theory and applications come together very conveniently under a Gaussian measure (P_G) as the well known Hermitian polynomials are not only orthogonal with respect to P_G but also they represent a basis of the space of $L^2(P_G)$ functions. This is unfortunately too much to ask when a different probability measure is selected leading to two further challenges, first the need to search of a set of orthogonal functions under the given measure P and secondly the shrinkage of the space $L^2(P)$ to a subspace in which the orthogonal functions become a basis. The latest have the strongest implications as it cut short the space of functions used for matching the marginal functions (conditional expectations) obtained from a regression between the dependent variable y and each of the independent variables X.

In this section we first explore changes on the setting developed in [1] in order to accommodate for non-Gaussian measures. In a second step a methodology to build a solution under a measure P is provided. This is based on the knowledge of

a basis of orthogonal functions, and their corresponding space Ω, under the fixed measure as well as a set of marginal functions, which are assumed elements of this space Ω. Then a solution for the case when the distribution between the factors is a multivariate t-student is developed. To accomplish that we describe a class of hypergeometric orthogonal polynomial which has the orthogonality relationship under the t-student measure. The space set up to solve the problem is therefore different from the Gaussian case.

In our approach the parameters of the chosen measure plays a role on the space under analysis and therefore on the set of suitable marginal functions. This is not the case for the Gaussian measure as the space is always that generated by the Hermitian polynomials regardless of the correlations, covariances of the underlying variables. This is why we also provide details of the problem for different degrees freedom starting from 4, 5 and 6 to conclude with a general solution.

3.1 Modified Setup

In this section we set up a new space which is a variation of that of ([1]) and suitable for our purpose in Theorem 3.4. The new space is a subspace of $L^2(p)$ such that we can obtain a set of orthogonal functions $\{p_m(X)\}_{m=1}^{Z}$ under a given measure P (i.e. $\int p_n(x)p_k(x)dP(x) = 0$ iff $n \neq k$) and therefore a basis for this subspace.

Let Ω_n be the space generated by, a possible infinite, set of orthogonal polynomials under a unidimensional probability measure P on variable X_n ($span\{(p_m(X_n))_{m=1}^{Z}\}$). Let us define $\Phi = \{(\phi_1, ..., \phi_N), \phi_n \in \Omega_n, E\phi_n(x_n) = 0 \ n = 1, ..., N\}$. Note $\Omega_n \subset L^2(P)$ so elements in Ω_n are L^2 integrable under the probability measure P. And finally the space of multivariate functions suitable to our purposes: $\Omega = H \bigcap L^2(P^{(N)})$, where $H = \left\{ \prod_{n=1}^{N} \phi_n(x_n) \mid (\phi_1, ..., \phi_N) \in \Phi \right\}$. In this new setting, the space Ω represents all functions L^2 integrable under a given N-dimensional measure $P^{(N)}$ that can be written as a product of elements in Ω_n.

Remark 3.1 *For clarity, we can think of the case where $P^{(N)}$ is a N-dimensional gaussian measure then Ω plays the role of all $L^2(P^{(N)})$ functions. This is due to the fact that the product of the spaces Ω_n generated by the Hermitian polynomials*

(denoted space H) form a basis of all multivariate polynomials and therefore of all $L^2(P^{(N)})$ functions ($H = L^2(P^{(N)})$).

Note also H plays an important role in our setting as it is basically the space of functions from which the optimal solution to problem (2) will be obtained. The richer this space this space the stronger and more general the result would be.

Next we extend Lemma 2.1 to show uniqueness of the solution to problem (2) on this modified space Ω. Recall that we use Pr_E to denote the orthogonal projection on a space E and $\|.\|$ the $L^2(P^{(N)})$ norm.

Lemma 3.2 *Assume $\hat{\phi}$ and ϕ satisfy (1) and are elements in Ω with*

$$\phi = \sum_{n=1}^{N} \psi_n(x_n)$$

then $\|\phi\| \leq \left\|\hat{\phi}\right\|$.

Proof 3.3 *The proof follows similarly to that of Lemma 2.1 but using a different space: Denote*

$$E_n = \{\xi \in \left(\Omega_n \bigcap L^2(P)\right) : \xi \quad is \quad X_n - measurable, E\xi = 0\}$$

Let us assume $\tilde{\phi}$ satisfies (1) , this means:

$$E(\tilde{\phi} \ |X_n) = \phi_n(X_n) = E(\phi|X_n)$$

or alternatively

$$Pr_{En}\tilde{\phi} = \phi_n = Pr_{En}\phi$$

which implies

$$\begin{aligned} Pr_{En}\tilde{\phi} - Pr_{En}\phi &= Pr_{En}(\tilde{\phi} - \phi) = 0 \\ n &= 1, ..., N \end{aligned}$$

Hence $(\tilde{\phi} - \phi)$ is orthogonal to E_n for $n = 1, ..., N$ and all $\xi_n \in E_n$:

$$\langle \tilde{\phi} - \phi, \xi_n \rangle = 0$$

As $\phi \in E_1 + E_2 + ... + E_N$ (which is the space of sums $\xi_1 + ... + \xi_N$ where $\xi_i \in E_i$) so $\phi = \xi_1 + ... + \xi_N$ so we can imply $\tilde{\phi} - \phi$ is orthogonal to ϕ using the linearity of expectation (scalar product):

$$\langle \tilde{\phi} - \phi, \phi \rangle = \langle \tilde{\phi} - \phi, \Sigma_{i=1}^{N} \xi_i \rangle = \Sigma_{i=1}^{N} \langle \tilde{\phi} - \phi, \xi_i \rangle = 0$$

This implies $||\tilde{\phi}|| \geq ||\phi||$, which follows easily from the Pythagorean theorem:

$$||\tilde{\phi}|| = ||\phi|| + ||(\phi - \tilde{\phi})||$$

and the equality is possible only if $\tilde{\phi} = \phi$.

Now we present the main theorem of this section which provides a methodology to build the minimum variance solution (problem (2)) given the following inputs:

- A marginal probability measure P and a set of orthogonal functions under this measure: $\{p_m(X)\}_{m=1}^{Z}$.
- A multivariate probability measure $P^{(N)}$ with marginals P.
- A set of marginal conditional expectations (1): $\{\varphi_n(X_n)\}_{n=1}^{N}$. These are obtained by, for example, marginal regressions (linear or nonlinear) between Y and each of the variables X_n. The functions $\varphi_n(X_n)$ must be in the space $\Omega_n = span\{(p_m(X_n))_{m=1}^{Z}\}$.

Theorem 3.4 *Let $P^{(N)}$ be a probability measure in $\Re^N$ with equal marginal P. Let $p_m(X)$, $m = 1, .., Z$ (may be infinite) be a set of orthogonal polynomial under the probability measure P. A set of functions $\varphi_n(x_n)$, $n = 1, .., N$ which are elements of Ω_n the vector space spanned by $\{p_m(X)\}_{m=1}^{Z}$ with coefficients $\{a_{nm}\}_{m=1}^{Z}$. Then the solution of (2) is given by the series:*

$$\phi(X_1, X_2, ..., X_N) = \sum_{n=1}^{N} \sum_{m=1}^{Z} \alpha_{nm} p_m(X_n) \quad , X_n \in \Re$$

whenever the series converges in $L^2(P^{(N)})$. Here α_{nm} are scalars solutions of:

$$a_{nm} = \sum_{n=1}^{N} \sum_{m=1}^{Z} \alpha_{ks} C_{kn}^{sm} \tag{4}$$

and C_{kn}^{sm} is defined in terms of a scalar product:

$$C_{kn}^{sm} = \langle p_s(X_k), p_m(X_n) \rangle$$

where

$$n, k = 1, ..., N \quad and \quad m, s = 1, ..., Z$$

Proof 3.5 *Thanks to our previous lemma, we only need to show that ϕ satisfies condition (1):*

$$E(\phi(x_1, x_2, ..., x_N \mid x_n) = \varphi_n(x_n) \quad n = 1, ..., N$$

consider the spaces E_n which is defined in proof 3.3.
Using the fact that each $\varphi_n(x_n)$ belongs Ω_n:

$$E[\phi|x_n] = Pr_{E_n}\phi = \sum_{m=1}^{Z} \langle \phi, p_m(x_n) \rangle p_m(x_n)$$

We next substitute our guess solution:

$$\begin{aligned} E[\phi|x_n] &= \sum_{m=1}^{Z} \langle \sum_{s=1}^{Z} \sum_{k=1}^{N} \alpha_{ks} p_s(x_k), p_m(x_n) \rangle p_m(x_n) \\ &= \sum_{m=1}^{Z} (\sum_{s=1}^{Z} \sum_{k=1}^{N} \alpha_{ks} \langle p_s(x_k), p_m(x_n) \rangle p_m(x_n)) \\ &= \sum_{m=1}^{Z} (\sum_{k,s=1}^{N,Z} \alpha_{ks} C_{kn}^{sm}) p_m(x_n) = \sum_{m=1}^{Z} a_{nm} p_m(x_n) \\ &= \varphi_n(x_n) \quad , n = 1, ..., N \end{aligned}$$

An application of Lemma (2.1) *completes the proof.*

Remark 3.6 *Note the meaning of the coefficients a_{nm} in our framework:*

$$a_{nm} = \langle \varphi_n, p_m(x_n) \rangle$$

therefore

$$\varphi_n(x_n) = \sum_{m=1}^{Z} a_{nm} p_m(x_n)$$

In order to gain better intuition about the solution and how it connects to the one found in [1] we show similarities between both approaches in the next section.

3.1.1 Gaussian and Independent Measures in our Framework

If we assume that P is a Gaussian distribution in Theorem 3.4 then $p_m(X)$ becomes the Hermitian polynomials $H_m(X)$, because the Hermite polynomials have an orthogonality relationship under a Gaussian measure and they are L^2 integrable. $H_n(x)$ nth-degree polynomials for $n = 0, 1, 2, 3, ...$ These polynomials are orthogonal with respect to the weight function (measure)

$$w(x) = \mathrm{e}^{-\mathrm{x}^2/2}$$

i.e., we have

$$\int_{-\infty}^{\infty} H_m(x) H_n(x)\, w(x)\, \mathrm{dx} = 0$$

when m is not equal to n. Furthermore,

$$\int_{-\infty}^{\infty} H_m(x) H_n(x)\, w(x)\, \mathrm{dx} = \sqrt{2\pi}\mathrm{n}!\delta_{\mathrm{nm}}$$

Therefore we have that $Z = \infty$ and the solution of (2) is given by the $L^2 -$ *convergent* series

$$\phi(X_1, ..., X_N) = \sum_{n,m=1}^{N,\infty} \alpha_{nm} H_m(X_n)$$

where

$$a_{nm} = \sum_{k=1}^{N} \alpha_{km} C_{kn}^m$$

and C_{kn}^m is defined, as before, in terms of a scalar product under the multivariate Gaussian measure:

$$C_{kn}^m = \langle H_m(X_k), H_m(X_n) \rangle$$

In this case $\langle H_m(X_k), H_k(X_n) \rangle = 0$ from Lemma $A.1$ from [1] simplifies the solution significantly.

Note that if the multivariate measure is an independent measure and therefore is the product of univariate probability distributions we will have:

$$\langle p_s(x_k), p_m(x_n)\rangle = 0 \;\text{ if }\; \{n \neq k\} \cap \{m \neq s\}$$

which leads to the following simplification:

$$\begin{aligned} a_{nm} &= \sum_{k,s=1}^{N,Z} \alpha_{ks} C_{kn}^{sm} \\ &= \langle \varphi_n(x_n), p_m(x_n)\rangle \end{aligned}$$

meaning the solution $\phi(X_1, X_2, ..., X_N)$ is:

$$\phi(X_1, X_2, ..., X_N) \;=\; \sum_{n1}^{N} \varphi_n(X_n),$$

which reproduces the findings in [1]. The advantage of reaching this same conclusion using Theorem 3.4 is that it becomes clear what type of functions would be best to use as $\varphi_n(X_n)$ candidates (those in Ω_n). If a set of orthogonal functions under P is known then we could build $\varphi_n(X_n)$ using elements in Ω_n in the regression analysis. On the other hand functions on the larger set $L^2(P)$ may be difficult to identify requiring trial and error on the type of functions $\varphi_n(X_n)$ to be used for the marginal regressions.

3.2 Results Under a t-Student Measure

This section focuses on an application of Theorem 3.4 that goes beyond the Gaussian probability measure and its multivariate generalization. We explore the case of the univariate t-student and one of the few bivariate t-student generalizations from [7], that allows for univariate t-student marginals. In general multivariate distributions may not give marginals from the same family, this is unfortunately the case of the t-student, and even though there are many proposed multivariate generalizations (see [9]) few of them have the standard univariate t-student as their marginal counterpart.

Some well known properties about the moments of the t-student are provided next. After that, a set of orthogonal polynomials under this measure is presented

together with the particularities about the orthogonality relationship and some implications from them.

Suppose that P is a t-student non degenerate distribution. We will assume that $E(X_n) = 0$ and v stands for the parameters, the degree of freedom with $v > 3$. The student's t-distribution has the probability density function

$$f(x) = (\frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi\Gamma(\frac{v}{2})}})(1 + \frac{x^2}{v})^{-\frac{1}{2}(1+v)}$$

The moments of the t-distribution are given as follow:

$$E(x^k) = \begin{cases} 0 & \text{if } k \text{ odd } ,0 < k < v \\ \frac{1}{\sqrt{\pi}\Gamma(v/2)}[\Gamma(\frac{k+1}{2})\Gamma(\frac{v-k}{2})v^{k/2} & \text{if } k \text{ even } ,0 < k < v \\ undefined & \text{if } k \text{ odd } ,0 < v \leq k \\ \infty & \text{if } k \text{ even } ,0 < v \leq k. \end{cases}$$

3.2.1 Orthogonal Polynomials and Properties

Orthogonal functions under a given probability measure is a key component of Theorem 3.4. Here we present a set of orthogonal polynomials under the univariate t-student distribution (see [14]). Consider the third finite class of classical hypergeometric orthogonal polynomials $I_n^p(x)$, $n = 0, 1, 2, ...$ One way to define them is as follows:

$$I_n^p(x) = n! \sum_{k=0}^{[\frac{n}{2}]} (-1)^k \begin{pmatrix} p-1 \\ n-k \end{pmatrix} \begin{pmatrix} n-k \\ k \end{pmatrix} (2x)^{n-2k}$$

which is equivalent to:

$$I_n^p(x) = \frac{(-2)^n (p-n)_n}{(2p-2n-1)_n} (1+x^2)^{p-1/2} \frac{d^n((1+x^2)^{n-(p-1/2)})}{dx^n} \quad n = 0, 1, 2, ...$$

where

$$(a)_n = a(a+1)(a+2)...(a+n-1)$$

A third way of defining them is as follows:

$$I_n^p(x) = \frac{1}{\sqrt{m!}} \frac{\partial^m}{\partial a^m}|_{a=0} (1 + 2ax - a^2)^{p-1}$$

Each definition provides a useful tool when manipulating these polynomials. The expression for the first 4 polynomials are:

$$\begin{aligned}
I_0^p(x) &= 1 \\
I_1^p(x) &= 2(p-1)x \\
I_2^p(x) &= 4(p-2)(p-1)x^2 - 2(p-1) \\
I_3^p(x) &= 8(p-3)(p-2)(p-1)x^3 - 12(p-2)(p-1)x
\end{aligned}$$

The subindex means the degree of the polynomial and the supraindex is the parameter of the measure in which they are orthogonal: $I_n^p(x)$ is nth-degree polynomial for $n = 0, 1, ...$ with parameter p. These polynomials are orthogonal with respect to the weight function

$$\rho(x,p) = (1+x^2)^{-(p-1/2)}.$$

In general we have the following orthogonality relationship:

$$\int_{-\infty}^{\infty} \rho(x,p) I_n^p(x) I_m^p(x) dx = (\frac{n! 2^{(2n-1)} \sqrt{\pi} \Gamma^2(p) \Gamma(2p-2n)}{(p-n-1)\Gamma(p-n)\Gamma(p-n+1/2)\Gamma(2p-n-1)}) \delta_{n,m}$$

This equation holds if and only if $m, n = 0, 1, 2, ..., N < p-1$ and $N = \max\{m, n\}$:

$$\int_{-\infty}^{\infty} \rho(x,p) I_n^p(x) I_m^p(x) dx = \begin{cases} 0 & \text{if } m = n \\ (\frac{n! 2^{(2n-1)} \sqrt{\pi} \Gamma^2(p) \Gamma(2p-2n)}{(p-n-1)\Gamma(p-n)\Gamma(p-n+1/2)\Gamma(2p-n-1)} & \text{if } m \neq n \end{cases}$$

In particular when $m = n$:

$$(\frac{(p-n-1)\Gamma(p-n)\Gamma(p-n+1/2)\Gamma(2p-n-1)}{n! 2^{(2n-1)} \sqrt{\pi} \Gamma^2(p) \Gamma(2p-2n)}) \int_{-\infty}^{\infty} \rho(x,p) (I_n^p(x))^2 dx = 1$$

which means the functions are not orthonormal but just orthogonal.

The weight function of the orthogonality relation above is related to the univariate t-student distribution as follows:

$T(x;v) = (\frac{\Gamma(v+1)/2}{\sqrt{v\pi}\Gamma(v/2)})\rho(\frac{x}{\sqrt{v}}; v/2+1)) \qquad -\infty < x < \infty.$

Therefore the orthogonality relationship in terms of a t-student with v degrees of freedom becomes:

$$\begin{aligned}
& \int_{-\infty}^{\infty} T(x;v) I_n^{v/2+1}(x) I_m^{v/2+1}(x) dx \\
= & \left(\frac{\Gamma(\frac{v+1}{2}) n! 2^{(2n-1)} \Gamma^2(v/2+1)\Gamma(v+2-2n)}{\Gamma(\frac{v}{2})(v/2-n)\Gamma(v/2+1-n)\Gamma(v/2-n+3/2)\Gamma(v-n+1)}\right)\delta_{n,m} \\
= & A(v,n)\delta_{n,m}
\end{aligned}$$

if and only if $m, n = 0, 1, 2, ..., Z < p-1$ therefore using the relation $p = v/2+1$ we have the following constraints for the number of orthogonal polynomials under this measure (Z)

$$Z = \begin{cases} \frac{v}{2} - 1 & \frac{v}{2} \in \mathbb{N} \\ [\frac{v}{2}] & \frac{v}{2} \notin \mathbb{N} \end{cases}$$

We could orthonormalize these polynomials by simply dividing by the function $A(v,n)$. The first four orthonormalized polynomials would look as follow:

$$I_0(x) = \sqrt{\left(\frac{(v/2)\Gamma(v/2+1)\Gamma(v/2+3/2)\Gamma(v+1)\Gamma(v/2)}{2!2^{(-1)}\Gamma((v+1)/2)\Gamma^2(v/2+1)\Gamma(v+2)}\right).1}$$

$$I_1(x) = \sqrt{\left(\frac{(v/2-1)\Gamma(v/2)\Gamma(v/2+1/2)\Gamma(v/2)}{2!\Gamma((v+1)/2)\Gamma^2(v/2+1)}\right).(v)}\frac{x}{\sqrt{(v)}}$$

$$I_2(x) = \sqrt{\left(\frac{(v/2-2)\Gamma(v/2-1)\Gamma(v/2-1/2)\Gamma(v-1)\Gamma(v/2)}{2!2^{(3)}\Gamma((v+1)/2)\Gamma^2(v/2+1)\Gamma(v-2)}\right.}.[(v-2)x^2 - v]$$

$$I_3(x) = \sqrt{\left(\frac{(v/2-3)\Gamma(v/2-2)\Gamma(v/2-3/2)\Gamma((v-2)\Gamma(v/2)}{3!2^{(5)}\Gamma((v+1)/2)\Gamma^2(v/2+1)\Gamma(v-4)}\right)}[(v-4)(v-2)\frac{x^3}{\sqrt{v}} - 3(v-2)(v)\frac{x}{\sqrt{v}}]$$

An important difference with the Gaussian case described before is the finiteness of the number of orthogonal polynomials under the t-student probability measure for a given degree of freedom v, this results from the upper limit for Z.

Remark 3.7 *For example, if $v = 4$ then $Z = 1$ so there are only two orthogonal polynomials under this t-student. Curiously if $v = 5$ or 6, then $Z = 2$ so we will have three orthogonal polynomials or those two cases (same number per case). On the other hand if the degrees of freedom increases to infinity then the number of polynomials also increases and not only the t-student becomes a Gaussian distribution but also the hypergeometric orthogonal polynomials converges to the Hermitian polynomials (see [14]).*

3.2.2 Multivariate t-Student

As mentioned before, most multivariate t-students are not built to imply univariate t-students on their marginals, this is why we are constrained to a very specific family of bivariate t-students which does have this condition, see [7]. Even though the chosen family has not been extended beyond two dimensions, we can still use it within the context of Theorem 3.4. The reason for this is that in order to apply Theorem 3.4, we only need to compute bivariate moments coming from the scalar products in matrix C_{kn}^{ms} $(\langle p_m(X_k), p_s(X_n)\rangle)$ and a bivariate distribution is all the requirement to do so. Still a multivariate distribution is needed for lemma 4.1 to ensure existence and uniqueness of the solution, so we make use of a theory still in its origins which blend univariate, bivariate and lower order dimensional distributions into a multivariate distribution of any dimension. This theory is that of Frechet classes ([8]).

Frechet classes or classes of multivariate distributions with some given marginals is one of the concepts that allows to combine the chosen bivariate into a N-dimensional multivariate distribution. The most popular Frechet classes are the classes of copulas. Copulas combine univariate distributions into a feasible multivariate probability measure. Another Frechet class is the class of multivariate distribution in which trivariate marginals are given. The idea is extended to any lower dimensional distribution. In our case we assume there is a multivariate t-distribution that can be obtained, using Frechet classes and properties, from a set of bivariate distribution.[8]. In other words, for the variables $(X_1, ..., X_N)$ and for a given set of bivariate distributions $\{P_{ij}(X_i, X_j)\}_{i,j=1}^{N}$, in this case the bivariate t-student [7], we assume a multivariate probability measure $P(X_1, ..., X_N)$ that can be built satisfying the given bivariate marginals. Conditions for this multivariate to exist are under study in the mathematical literature as it remains an open problem. The only fully solve case is when the marginals are univariate (the case of Copulas mentioned before).

Remark 3.8 *It should be pointed out that there are sufficient conditions available for the existence of N-dimensional probability distributions from a given set of 2-dimensional marginals. This, in a worst case scenario, would only mean additional constraints on the parametric space for the selected bivariate distributions.*

The next remark is about the implications that result from using only bivariate to the space of functions H from which the optimal function solving problem (2) is obtained.

Remark 3.9 *Note that in a worst case scenario, where the implied multivariate distribution has few $L^2(P^{(N)})$ integrable elements we would still have functions of the type $\{p_m(X_k)p_s(X_n)\}_{m,s=1}^{Z}$ for all $n, k = 1, .., N$ due to the existence of the joint moments from our chosen bivariate marginal to be described later.*

The bivariate t-distribution we selected, was derived in [7] as a sampling distribution from the bivariate normal distribution and chi-square distribution. The joint probability density function is given by the following expression:

$$f(X_1, X_2; v, \rho) = \frac{(1-\rho^2)}{2\pi}(1 + \frac{X_1^2 + X_2^2 - 2\rho X_1 X_2}{(1-\rho^2)^v})^{-(v+2)/2}$$

where v represents the degrees of freedom in the one dimensional marginals and ρ is the dependence parameter with happens to coincide with the Pearson correlation between the variables.

Therefore, in principle we assume that the joint probability measure $P(X_1, ..., X_N)$ is such that every bivariate (X_i, X_j) with $i, j = 1, ..., N$ follows a two-dimensional t-student with parameters (v, ρ_{ij}).

The joint moments of this bivariate t-distribution are given in general by a recursive system as:

$$\mu(a, b; v) = (a+b-1)\rho\mu(a-1, b-1)\gamma_2 + (a-1)(b-1)(1-\rho^2)\mu(a-2, b-2)\gamma_4 \text{ if } v > 4$$

For even and odd joint moments the expression can be simplified to:

$$\begin{cases} \mu(2a, 2b; v) = \frac{(2a)!(2b)!}{2^{a+b}} \sum_{j=0}^{min(a,b)} \frac{(2\rho)^{2j}}{(a-j)!(b-j)!(2j)!}\gamma_{2a+2b} & \text{if } v > 2a+2b \\ \\ \mu(2a+1, 2b+1; v) = \frac{(2a+1)!(2b+1)!}{2^{a+b}}\rho \sum_{j=0}^{min(a,b)} \frac{(2\rho)^{2j}}{(a-j)!(b-j)!(2j+1)!}\gamma_{2a+2b+2} & \text{if } v > 2a+2b+2 \\ \\ \mu(2a, 2b+1; v) = \mu(2a+1, 2b; v) = 0, & \end{cases}$$

where $\mu(a,b;v) = E(x_1^a x_2^b)$, and $\sigma_1 = \sigma_2 = 1$ and γ_a is the a-th moment of new distribution defined by the authors as "T" (see [7], Theorem in page 7).

T is a random variable with an inverted chi-square distribution given by Theorem 1.1 in [7]. The a-th moment of T is given by:

$$\begin{cases} \gamma_a = E(T^a) = \frac{(v/2)^{a/2}\Gamma(v/2-a/2)}{\Gamma(v/2)} \quad \text{for } v > a \\ \\ \gamma_{-2a} = E(T^{-2a}) = (v/2)^{-a}(v/2)(v/2+1)\cdots(v/2+a-1) \quad \text{if } v > 2a \\ \\ \gamma_{2a} = E(T^{2a}) = \frac{(v/2)^a}{(v/2-1)(v/2-2)\cdots(v/2-a)} \quad \text{for } v > 2a \end{cases}$$

Note $\gamma_2 = \frac{v}{v-2}$ and therefore $v > 2$ (see [7] Corollary 1.1, page 2).

In order to provide a bit more of insight regarding this bivariate t-distribution we will show the connection to the bivariate Gaussian distribution. Recall the raw product moments of the bivariate normal distribution with pdf:

$$\begin{aligned} f(x_1,x_2) &= \frac{(1-\rho)^{-1/2}}{2\pi\sigma_1\sigma_2} exp(\frac{-q(x_1,x_2)}{2}) \\ (1-\rho^2)q(x_1,x_2) &= (\frac{x_1-\mu_1}{\sigma_1} + \frac{x_2-\mu_2}{\sigma_2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}) \end{aligned}$$

are given by

$$\mu(a,b) = \sigma_1^a \sigma_2^b \lambda(a,b)$$

where in general there is a recursive expression for $\lambda(a,b)$:

$$\lambda(a,b) = (a+b-1)\rho\lambda(a-1,b-1) + (a-1,b-1)(1-\rho^2)\sigma_1^2\sigma_2^2\mu(a-2,b-2)$$

which is solvable for odd and even joint moments as:

$$\begin{cases} \lambda(2a,2b) = \frac{(2a)!(2b)!}{2^{a+b}} \sum_{j=0}^{min(a,b)} \frac{(2\rho)^{2j}}{(a-j)!(b-j)!(2j)!} \\ \lambda(2a+1,2b+1) = \frac{(2a+1)!(2b+1)!}{2^{a+b}} \rho \sum_{j=0}^{min(a,b)} \frac{(2\rho)^{2j}}{(a-j)!(b-j)!(2j+1)!} \\ \lambda(2a,2b+1) = \lambda(2a+1,2b) = 0 \end{cases}$$

(see [7] page 2, Theorem 3.1)

These equations help us realize that when $v \to \infty$, the pdf of the bivariate t-distribution converges to the one of a bivariate normal distribution. This is a plus for the chosen bivariate t-student selected as it is a natural extension not only of a univariate gaussian but also of its bivariate counterpart.

Note that in setting the stage for the application of Theorem 3.4 to the t-student case, we also need the scalar product of the orthogonal functions under a bivariate t-student probability measure, this is the matrix $C_{kn}^{ms} = \langle p_m(X_k), p_s(X_n) \rangle$. The next result gives the values for this matrix:

Lemma 3.10 *The inner product with respect to the joint t-student of the orthogonal polynomials is:*

$$\begin{aligned} C_{kn}^{ms} &= \langle I_s^p(x_k), I_m^p(x_n) \rangle \\ &= m!s! \sum_{a=0}^{[m/2]} \sum_{b=0}^{[s/2]} (-1)^{a+b} 2^{s+m-2a-2b} \mu(s-2b, m-2a; v) \tbinom{p-1}{s-b} \tbinom{s-b}{b} \tbinom{p-1}{m-a} \tbinom{m-a}{a} \end{aligned}$$

Where $\mu(a,b;v)$ was previously defined.

Proof 3.11 *The proof follows easily from the joint moments under the bivariate t-student distribution.*

At this point we have all the inputs needed to use Theorem 3.4 and therefore to present the solution for the t-student case. Still due to the richness of cases implied by the dependence of the solution to the degrees of freedom we will then developed three particular cases as a first step in the next section.

3.2.3 Solution for a t-Student. Three Particular Cases

We started with 4 degrees of freedom which is the most nongaussian case of all because when $v <= 3$ then $Z = 0$ and there is no basis for the space so $v => 4$ and when $v \to \infty$, the pdf of the bivariate t-distribution converges to the one of a bivariate normal distribution. We provide the results for the case of 5 and 6 degrees of freedom. The reason behind developing both the $v = 5$ and the $v = 6$ is to show that even though the numbers of polynomials are the same we still obtain different solutions to our main problem (2) from Theorem 3.4.

Solution for v=4 Let us assume the N variables $(X_1, .., X_N)$ follow a multivariate t-student distribution such that every pair (X_i, X_j) follows a bivariate t-student as defined previously with parameters (v, ρ_{ij}).

When the degree of freedom is 4 we have $\frac{v}{2} = 2$ so $Z = v/2 - 1 = 1$. The condition of zero expected value for the input functions $\varphi_n(X)$ implies that we do not need or use the orthogonal function $I_0 = 1$.

Therefore I_1 is the only orthogonal basis in the space E_n, the vector solution from Theorem 3.4 would then be:

$$\begin{pmatrix} \alpha_{11} \\ \vdots \\ \alpha_{N1} \end{pmatrix} = C^{-1} \begin{pmatrix} a_{11} \\ \vdots \\ a_{N1} \end{pmatrix}$$

where

$$\begin{aligned} a_{n1} &= E(\varphi_n(X_n) I_1(X_n)) \\ &= \int_{\Re} \varphi_n(X) I_1(X) dP(X) \quad n = 1, 2, ..., N. \end{aligned}$$

and $P(X)$ is the univariate t-student measure.

The degree of φ_n must be 1 because $\varphi_n \in span\{I_1\}$ and I_1 is a polynomial of degree one. The matrix C is as follows:

$$C(i, j) = \left(\ \langle I_1(x_i), I_1(x_j) \rangle \ \right)$$

which corresponds to

$$C = a.R$$

$$R = \begin{pmatrix} 1 & \rho_{21} & ... & \rho_{N1} \\ \rho_{12} & 1 & ... & \rho_{N2} \\ \vdots & & & \\ \rho_{1N} & \rho_{2N} & ... & 1 \\ & & & \end{pmatrix}$$

a is a constant and R is the correlation matrix. The determinant of the correlation matrix will equal 1 only if all correlations equal 0, otherwise the determinant will be less than 1. It is not zero either since the factors are different $X_i \neq X_j$ when $i, j = 1, ..., N$ (non degenerated distribution) so C is invertible therefore the vector α is well defined. The solution of (2) is given by

$$\phi(X_1, X_2, ..., X_N) = \sum_{n=1}^{N} \alpha_{n1} I_1(X_n) \quad , X_n \in \Re$$

Solution for v=5 We uses the same setting as in the case $v = 4$. Here when degree of freedom is 5 then we have $\frac{v}{2} = 2.5$ so $Z = [2.5] = 2$ therefore I_1 and I_2 are the orthogonal elements in the space E_n. Using Theorem 3.4, the vector solution would be:

$$\begin{pmatrix} \alpha_{11} \\ \vdots \\ \alpha_{N1} \\ \alpha_{12} \\ \vdots \\ \alpha_{N2} \end{pmatrix} = C^{-1} \begin{pmatrix} a_{11} \\ \vdots \\ a_{N1} \\ a_{12} \\ \vdots \\ a_{N2} \end{pmatrix}$$

where

$$\begin{aligned} a_{nm} &= E(\varphi_n(X_n) I_m(X_n)) \\ &= \int_{\Re} \varphi_n(X) I_m(X) f(X) dX \\ & n = 1, 2, ..., N \quad m = 1, 2 \end{aligned}$$

and $f(X)$ is the probability density function of the t-student:

$$f(x) = (\frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi\Gamma(\frac{v}{2})}})(1+\frac{x^2}{v})^{-\frac{1}{2}(1+v)}$$

The degree of polynomial φ_n can be 1 or 2 because $\varphi_n \in span\{I_1, I_2\}$, I_1 is a polynomial of degree one and I_2 is a polynomial of degree two.

From the orthogonality relationship of hypergeometric polynomials we obtained the following form for C:

$$C_{ji}^{mn} = \langle I_m(x_j), I_n(x_i)\rangle = \begin{cases} 0 & \text{if } m \neq n \\ \langle I_m(x_j), I_m(x_i)\rangle & \text{if } m = n \end{cases}$$

which can be simplified further and C becomes a block diagonal matrix:

$$C = \begin{pmatrix} C_{11}^{11} & . & C_{N1}^{11} & 0 & . & 0 \\ . & & & & & \\ C_{1N}^{11} & . & C_{NN}^{11} & 0 & . & 0 \\ 0 & 0 & 0 & C_{11}^{22} & . & C_{N1}^{22} \\ . & & & & & \\ 0 & 0 & 0 & C_{1N}^{22} & . & C_{NN}^{22} \end{pmatrix}$$

Therefore for each $m = 1, 2$ we can write:

$$\begin{pmatrix} \alpha_{1m} \\ \vdots \\ \alpha_{Nm} \end{pmatrix} = B_m^{-1} \begin{pmatrix} a_{1m} \\ \vdots \\ a_{Nm} \end{pmatrix}$$

where

$$B_m(i,j) = C_{ji}^{mn} = \left(\ \langle I_m(x_i), I_m(x_i)\rangle \ \right)$$

In the case of $m = 1$, we have:

$$B_1 = b.R$$

where b is a constant and R is the correlation matrix hence B_1 is invertible. For

$m = 2$:

$$B_2 = \begin{pmatrix} 22 & 7+15\rho_{21}^2 & ... & 7+15\rho_{N1}^2 \\ 7+15\rho_{12}^2 & 22 & ... & 7+15\rho_{N2}^2 \\ \vdots & & & \\ 7+15\rho_{1N}^2 & 7+15\rho_{2N}^2 & ... & 22 \end{pmatrix}$$

Proposition 3.12 *The determinant of the matrix B_2 is nonzero if the matrix $\Sigma = (\rho_{ij})$ is nondegenerated and $\sum_{j=1,\neq i}^{N} \rho_{ij} < 1$ for all i.*

Proof 3.13 *Note the determinant of the matrix is B_2 zero if $\rho_{21}^2 = ... = \rho_{NN-1}^2 = 1$ which is not possible (the factors are not equal) but this is only a sufficient condition.*

In general B_2 can be written as $B_2 = aO + b\Sigma^*$ where a and b are scalars (7 and 15), O is a matrix of ones and Σ^* is a nondegenerated covariance matrix, this last statement comes from using $\Sigma = (\rho_{ij})$ is nondegenerated and realizing that a matrix with square correlations can be obtained using factor analysis (see [2]) as follows (for simplicity we show only in dimension 3):

$$\begin{cases} W_1 = \rho_{12}M_1 + \rho_{13}M_2 + \sqrt{1-\rho_{12}^2-\rho_{13}^2}Z_1 \\ W_2 = \rho_{12}M_1 + \rho_{23}M_3 + \sqrt{1-\rho_{12}^2-\rho_{23}^2}Z_2 \\ W_3 = \rho_{13}M_2 + \rho_{23}M_3 + \sqrt{1-\rho_{13}^2-\rho_{23}^2}Z_3 \end{cases}$$

where M and Z are uncorrelated therefore the resulting covariance matrix is well defined (the argument in the square roots are positive) and nondegenerated. Moreover B_2 can be interpreted as representing the covariance matrix of a vector Z^* in a separate factor model $Z^*_{Nx1} = aM^*_{1x1} + bW_{Nx1}$, where M^* and W are independent with unit variance and W has covariance matrix Σ, factors models lead to nondegenerated matrixes (see [2]). Therefore B_2 is invertible.

The solution from Theorem 3.4 and $v = 5$ would be:

$$\phi(X_1, X_2, ..., X_N) = \sum_{n,m=1}^{N,2} \alpha_{nm} I_m(X_n) \quad , X_n \in \Re$$

Solution for v=6 Using the same setting as before, the solution for $v = 6$ is similar to the case of $v = 5$. When the degree of freedom is 6 then we have $\frac{v}{2} = 3$ so $Z = 2$ (same as for $v = 5$). The vectors I_1 and I_2 are the only elements in the orthogonal basis of E_n and the vector solution would be:

$$\begin{pmatrix} \alpha_{11} \\ \vdots \\ \alpha_{N1} \\ \alpha_{12} \\ \vdots \\ \alpha_{N2} \end{pmatrix} = C^{-1} \begin{pmatrix} a_{11} \\ \vdots \\ a_{N1} \\ a_{12} \\ \vdots \\ a_{N2} \end{pmatrix}$$

The degree of polynomial φ_n is again either 1 or 2 and C has a similar form as in the case $v = 5$:

$$C_{ji}^{mn} = \langle I_m(x_j), I_n(x_i) \rangle = \begin{cases} 0 & \text{if } m \neq n \\ \langle I_m(x_j), I_m(x_i) \rangle & \text{if } m = n \end{cases}$$

which can be written as a block diagonal. Therefore for each $m = 1, 2$ we can state:

$$\begin{pmatrix} \alpha_{1m} \\ \vdots \\ \alpha_{Nm} \end{pmatrix} = B_m^{-1} \begin{pmatrix} a_{1m} \\ \vdots \\ a_{Nm} \end{pmatrix}$$

where

$$B_m(i, j) = \big(\ \langle I_m(x_i), I_m(x_j) \rangle \ \big)$$

In particular, for $m = 1$ we have the same expression as for $v = 5$:

$$B_1 = b.R$$

so B_1 is invertible. But for $m = 2$ the solution differs from that found in the $v = 5$ case:

$$B_2 = \begin{pmatrix} 204 & 60 + 144\rho_{21}^2 & ... & 60 + 144\rho_{N1}^2 \\ 60 + 144\rho_{12}^2 & 204 & ... & 60 + 144\rho_{N2}^2 \\ \vdots & & & \\ 60 + 144\rho_{1N}^2 & 60 + 144\rho_{2N}^2 & ... & 204 \\ & & & \end{pmatrix}$$

The reason for this is that scalar products depend on the probability measure, under a measure with 6 degrees of freedom are different that under a measure with 5 degrees of freedom. B_2 is still invertible using similar arguments as in the case $v = 5$. The solution of (2) using Theorem 3.4 for $v = 6$ is

$$\phi(X_1, X_2, ..., X_N) = \sum_{n,m=1}^{N,2} \alpha_{nm} I_m(X_n) \quad , X_n \in \Re$$

Once again note that even though the solutions look similar for $v = 5$ and $v = 6$ they do not give exactly the same optimal solutions.

3.2.4 Solution for a t-Student. General Case

In this section we target the solution for any degrees of freedom of the multivariate t-student introduced before. The key condition in order to reach a solution based on Theorem 3.4 is invertibility of the matrix C_{ji}^{mn} $(\langle I_m(x_j), I_n(x_i)\rangle)$. If this is proved then the values α can be obtained from a given set of a's.

Let us write C as a column matrix $C = (a_1 a_2 ... a_Q)$ therefore a_i is defined as:

$$a_i' = \begin{pmatrix} E(y_i y_1) & ... & E(y_i y_Q) \end{pmatrix}$$

Here $E(y_i y_j) = \int\int y_i y_j f(x_l, x_m) dx_l dx_m$ where $f(x_l, x_m)$ is the probability density function of the 2-dimensional t-student distribution of x_l and x_m, $l, m = 1, .., N$. In our case the vector y is related to the Z orthogonal polynomials and N variables as follows $(Q = Z \cdot N)$

$$\begin{aligned} y &= (y_1, y_2, ..., y_Q) \\ &= (I_1(X_1), ..., I_1(X_N), I_2(X_1), ..., I_2(X_N), ..., I_z(X_1), ..., I_z(X_N)) \end{aligned}$$

In the next theorem we show that the invertibility of C can be implied from the linear independence of the y.

Theorem 3.14 *The determinant of matrix C is zero iff one of the y is a linear combination of the other $y's$.*

Proof 3.15 *Without losing generality we will show the proof for the case $y_Q = \Sigma_{i=1}^{Q-1} b_i y_i$. ($\Leftarrow$)*
From

$$y_Q = \Sigma_{i=1}^{Q-1} b_i y_i$$

we can says that:

$$E_{p^N}[y.y_Q] = \Sigma_{i=1}^{Q-1} b_i E_{p^N}[y.y_i]$$

or

$$E(y_Q y_j) = \Sigma_{i=1}^{Q-1} b_i E(y_i y_j) \quad for \quad j = 1, ..., Q$$

Hence,

$$\begin{pmatrix} E(y_Q y_1) \\ . \\ E(y_Q y_Q) \end{pmatrix} = b_1 \begin{pmatrix} E(y_1 y_1) \\ . \\ E(y_1 y_Q) \end{pmatrix} + b_2 \begin{pmatrix} E(y_2 y_1) \\ . \\ E(y_2 y_Q) \end{pmatrix} + ...$$

Therefore $a_Q = \Sigma_{i=1}^{Q-1} b_i a_i$ since one column of matrix C is a linear combination of the other columns of matrix C the determinant of C is zero.
($\Rightarrow$)

$$detC = 0$$

It means that one of the column of matrix C is a linear combination of other columns of matrix C. Therefore

$$a_n = \Sigma_{i=1}^{Q-1} b_i a_i \quad with \quad b_i \neq 0 \quad for \quad i = 1, ..., Q-1$$

Hence

$$E(y_k y_Q) = \Sigma_{i=1}^{Q-1} b_i E(y_k y_i) \quad for \quad k = 1, ..., Q$$

This implies:

$$\Sigma_{i=1}^{Q-1} b_i E(y_k y_i) - E(y_k y_Q) = E(y_k(\Sigma_{i=1}^{Q-1} b_i y_i - y_Q))) = 0 \quad for \quad k = 1, ..., Q$$

Now let $z = (\Sigma_{i=1}^{Q-1} b_i y_i) - y_Q = \Sigma_{i=1}^{Q} b_i y_i$ when $b_Q = -1$ then

$$E(y_k(\Sigma_{i=1}^{Q-1} b_i y_i - y_Q))) = E(y_k z) = 0 \quad for \quad k = 1, ..., Q$$

$$0 = \int ... \int y_k z f(x_1, ..., x_N) dx_1 ... dx_N, k = 1, ..., Q$$

Hence

$$\int ... \int z^2 f(x_1, ..., x_N) dx_1 ... dx_N = 0$$

As a result

$$||z||_2 = 0$$

which implies:

$$y_Q = \Sigma_{i=1}^{Q-1} b_i y_i.$$

The next remark shows that the previous result holds for the matrix C in our framework.

Remark 3.16 *As the hypergeometric polynomials are linearly independent then none of the polynomials for a fixed variable x_j can be written as a linear combination of the remaining polynomials, for simplicity:*

$$I_Z(x_j) \neq \sum_{i=1}^{Z-1} c_i I_i(x_j)$$

with $j = 1, .., N$. This is obvious due to the degree of the polynomials. Note moreover that the variables themselves are linearly independent from each other for any power (for simplicity $I_i(x_N) \neq \sum_{j=1}^{N-1} b_j I_i(x_j)$ for any i), this comes from assuming a non singular covariance matrix for the variables x and therefore a non-degenerated multivariate probability measure for the vector $(x_1, ..., x_N)$. Then we can conclude that C is invertible.

The application of Theorem 3.4 to the case of the t-student is presented in the next corollary.

Corollary 3.16.1 *The solution of (2) for the case when the distribution between the x (financial factors) is the joint t-student and $(I_m(X_n)_{m=1}^{Z}$ forms an orthogonal basis in Ω_n then the solution has the form:*

$$\phi(X_1, X_2, ..., X_N) = \sum_{n,m=1}^{N,Z} \alpha_{nm} I_m(X_n) \quad , X_n \in \Re$$

where α_{nm} is the solution of

$$a_{nm} = \sum_{k,s=1}^{N,Z} \alpha_{ks} C_{kn}^{sm}$$

$$Z = \begin{cases} \frac{v}{2} - 1 & \frac{v}{2} \in \mathbb{N} \\ [\frac{v}{2}] & \frac{v}{2} \notin \mathbb{N} \end{cases}$$

$$a_{nm} = E(\varphi_n(X_n) I_m(X_n)) = \int_{\Re} \varphi_n(X) I_m(X) f(X) dX$$

$$n = 1, 2, ..., N \quad m = 1, 2, ..., Z$$

and $\varphi_n \in span\{I_1, I_2, ..., I_Z\}$.

Some details about the solution are provided next: Let us rewrite matrix C (a square ZN matrix) as follows:

$$C = \begin{pmatrix} B_{11} & B_{12} & . & B_{1Z} \\ B_{21} & B_{22} & . & B_{2Z} \\ B_{31} & B_{32} & . & B_{3Z} \\ . & & & \\ B_{Z1} & B_{Z2} & . & B_{ZZ} \end{pmatrix}$$

where each B_{ij} is a $N \times N$ matrix such that:

$$B_{ij} = \begin{pmatrix} C_{11}^{ij} & C_{21}^{ij} & . & C_{N1}^{ij} \\ C_{12}^{ij} & C_{22}^{ij} & . & C_{N2}^{ij} \\ . & & & \\ C_{1N}^{ij} & C_{2N}^{ij} & . & C_{NN}^{ij} \end{pmatrix}$$

therefore

$$B_{ij}(m, n) = \langle I_i(x_n), I_j(x_m) \rangle, \quad m, n = 1, 2, ..., N$$

Note that $\langle I_i(x_n), I_j(x_m) \rangle = 0$ if $i + j$ is odd $(\mu(1, 0; v) = \mu(1, 2; v) = ... = \mu(i, j; v) = 0$ when $i + j = odd)$ so matrix C becomes:

$$C = \begin{pmatrix} B_{11} & 0 & B_{13} & 0 & . \\ 0 & B_{22} & 0 & B_{24} & . \\ B_{31} & 0 & B_{33} & 0 & . \\ . & & & & \end{pmatrix}$$

C is invertible therefore the vector solution of α would then be defined as:

$$\begin{pmatrix} \alpha_{11} \\ . \\ \alpha_{1Z} \\ \alpha_{2Z} \\ . \\ \alpha_{NZ} \end{pmatrix} = C^{-1} \begin{pmatrix} a_{11} \\ . \\ a_{1Z} \\ a_{2Z} \\ . \\ a_{NZ} \end{pmatrix}$$

so the solution of (2) for any v is:

$$\phi(X_1, X_2, ..., X_N) = \sum_{n,m=1}^{N,Z} \alpha_{nm} I_m(X_n) \quad , X_n \in \Re.$$

3.3 Other Probability Measures

The analysis presented in this chapter could be extended to other distributions and orthogonal functions. Some examples are given below:

Laguerre Polynomials (L_n) under an exponential probability measure (P). Here L_n is the solution of a second-order linear differential equation:

$$\begin{aligned} xL_n''(x) + (1-x)L_n'(x) + nL_n(x) &= 0 \\ P(x) &= e^{-x} \end{aligned}$$

There are many multivariate generalizations of the exponential distribution, see [9] so there is plenty of room for exploring our setting under this measure.

Associated Laguerre Polynomials (LA_n) under a Gamma probability measure (P). Here LA_n is also the solution of a second-order linear differential equation:

$$\begin{aligned} xLA_n''(x) + (\alpha+1-x)LA_n'(x) + nLA_n(x) &= 0 \\ P(x) &= x^{\alpha}e^{-x} \end{aligned}$$

There are only few multivariate variants of the Gamma distribution (see [9]) which could be worth exploring under our setting.

These last two cases are defined in $(0, \infty)$ so they apply better to positive relationships between financial variables as those observed between stock prices and index values.

4 Conclusions

The work in this thesis tackles an important problem in financial mathematics, which is that of explaining a given variable, like a stock price, using a vector of other financial/economical variables like Fundamentals, Indexes or Factors under few observations. Finding relationships between a single dependent variable and a set of independent variables is a standard problem of Regression analysis. It is well known that Regression leads to reliable results in the presence of medium to large sample sizes. Unfortunately there are several cases in applications were the data available is only of few dozen observations and the number of variables of interest is almost 50% of the number of observations, this is the case of Hedgefund data.

A recently published paper, [1], tried to overcome this lack of data and the shortfalls of regression by providing an alternative method that uses the joint distribution of the independent variables ($P^{(N)}$) on a kind of non-parametric construction of the best fit function between the dependent and indepedendent variables. This optimal fit function is obtained as the one with minimum overall variance among all L^2 integrable N-dimensional functions under the given multivariate probability measure $P^{(N)}$. His work is applied to the case of an independent measure and the multivariate Gaussian measure.

Our work focuses on the shortfalls of [1] and provides two main outcomes: it first makes the results in [1] more flexible by slightly modifying the type of functions on which the optimality is found. This modification allows for an explicit construction of the optimal solution based on a given univariate probability measure, a set of orthogonal functions under this measure and a multivariate probability measure that has given univariate as marginals. The second component of our thesis is the application of the first result on a specific measure, the t-student distribution. This distribution has fat tails and tail dependence, therefore it represents a more realistic probability measure for economic/financial factors. These two contributions could be used mainly in the context of risk management (as [1]), in particular it should approximate better the true relationship between returns and economical factors and therefore allow for building hedge portfolios closer to the targeted assets. The alternative approach of nonlinear regression should lead to unreliable estimators due to small sample sizes making our suggestion more appealing to practitioners.

5 References

References

[1] Cherney, Alexander, Raphael Douady, and Stanislav Molchanov. "Measuring Risk With Scarce Observation." (March 2008).

[2] Andrew Laurence Comrey, Howard Bing Lee. 1992 A first course in factor analysis. Taylor and Francis.

[3] Fox, John. "Nonlinear Regression And Nonlinear Least Square. "(January 2002)

[4] Goodman, David J. "Probability and Stochastic Processes."Second Edition. John Wiley and Sons, Inc (2005).

[5] Ho-nam, Eric. "Simple linear regression. "lecture 37.

[6] "Hedge Fund Databases" EUREKAHEDGE. copyright 2011 Eurekahedge Pte Ltd.

[7] Joarder, Anwar H. "Moments Of the T-Distribution." (2006).

[8] Joe, Harry."Multivariate Models and Dependence Concepts."(1997).

[9] Samuel Kotz, N. Balakrishnan, Norman L. Johnson (2000). Continuous Multivariate Distributions, Models and Applications, 2nd Edition.

[10] Kahane, Leo H. " Regression Basics." copyright 2001 by sage publications inc

[11] Kleinbaum, Kupper."Applied Regression Analysis And Other Multivariable Methods." Boston , Massachusetts.(1978)

[12] Lawson, C.L ;Hanson, R.J. (1974). Solving least Square Problems. Englewood Cliffs, NJ:Prentice-Hall

[13] Long, Scott. "Regression Models for Categorical and Limited Dependent Variables."(1997)

[14] Masjedjamei, Mohammad. "Three Finite Classes Of Hypergeometric Orthogonal Polynomials And Their Application In Functions Approximation. " Integral Transforms and Special Functions, 2002, Vol.13, pp.169-190.

[15] McCrary, Stuart A. "Hedge Fund A Professional's Guide."(2002)

[16] Samprit Chatterjee, Ali S.Hadi. "Regression Analysis By Example."Hoboken, N.J :Wiley-Interscience, 4th ed, c2006.

[17] Shaw, W.T, K.T.A. Lee. "Copula Method Vs Canonical Multivariate Distributions The Multivatiate Student T Distribution With General Degree Of Freedom." (2007).

[18] Sosa, Walter. "Conditional Expectation and linear regression."(2009)

[19] Van Huffel, Sabine. Vandewalle, Joos. "The total least square Problems".(1991)

[20] Weisberg, Sanford. "Applied Linear Regression."Third edition. (2005)

[21] Wonnacott, Thomas. H. "Regression:A Second Course In Statistics." (1981)