Abstract

This paper considers the problem of predicting an outcome variable using high-dimensional data. To control the overfit arising from the high-dimensionality one can use dimension reduction methods, which try reduce the set of predictors to a small set of orthogonal linear combinations of the predictor variables, which are subsequently used to predict the outcome. Examples are principal components regression and partial least squares. These methods are usually not motivated by a model and have strictly separated dimension reduction and prediction steps.

This paper looks at dimension reduction for high-dimensional regression from a modelling point of view. We propose a very general model for the joint distribution of the outcome and the predictor variable. This model is based only on the assumption that a set of latent variables exists such that outcome and the predictor variables are conditionally independent given the latent variables. We do not assume that the number of latent variables is known and we allow a very general error structure in the predictor variables. This model allows us to study the dimension reduction and prediction steps jointly.

In this model, we study parameter estimation and prediction in the situation where the number of predictor variables goes to infinity, while the number of samples remains fixed. Based on this analysis, we argue for a doing principal components regression with a relatively small number of components and using only a subset of the predictor variables, selected for their correlation with the outcome variable. This is a variant of the supervised principal components method proposed by Bair et al. on the basis of a much more restrictive model.

This chapter has been submitted as: J. J. Goeman and J. C. van Houwelingen. Model-based dimension reduction for high-dimensional regression.
6.1 Introduction

In recent years high-dimensional data have become increasingly common in many fields of science. This has attracted the attention of the statistical community, resulting in a surge of novel and interesting methodology.

In this paper we consider the basic high-dimensional prediction problem of predicting an outcome $y$ from a vector $x = (x_1, \ldots, x_p)'$ of predictors. The goal is to predict a new observation $y_{\text{new}}$ from an observed data vector $x_{\text{new}}$. The prediction rule for predicting $y_{\text{new}}$ from $x_{\text{new}}$ is to be constructed using a training sample of size $n$ from the joint distribution of $x_1, \ldots, x_p$ and $y$. The training data are gathered in an $n \times p$ data matrix $X$ and an $n \times 1$ vector $y$. The prediction problem is high-dimensional when the number of predictors $p$ is very large, typically larger than the size $n$ of the training sample. The overfit arising from the high number of predictors makes most classical statistical methods unusable.

Many different strategies have been proposed to counter the problem of overfit due to high dimensionality (Hastie et al., 2001). For example, variable selection methods reduce the dimensionality directly by selecting a subset of the predictors to be used for prediction. Shrinkage methods restrain the parameter estimates to prevent overfit (Hoerl and Kennard, 1970; Tibshirani, 1996; Van Houwelingen, 2001). Dimension reduction methods reduce the dimensionality of the prediction problem by using only a small number of orthogonal linear combinations of the original predictor variables (Jolliffe, 2002; Wold et al., 1984). All methods that control overfit in high-dimensional prediction share the property that they reduce the variance of the prediction while introducing bias. The methods differ mainly in the kind of bias introduced.

There is not one strategy or method that is known to be overall superior to all the others. As each method introduces bias, it will tend to perform well especially when the bias introduced is bias ‘towards the truth’. For example, a variable selection method can be expected to work best when most of the true regression coefficients are virtually zero. A ridge regression (Hoerl and Kennard, 1970) would most likely work well when most true regression coefficients are in reality small, but not zero. The choice of the method should therefore depend on knowledge or ideas about the underlying ‘truth’. Notions about the true relationships between the predictor variables and the outcome can help determine which method is best for which type of data. For a major part, the choice of method should be a modelling issue.

In this paper we investigate high-dimensional regression from a modelling perspective. We formulate a very general model for the joint distribution of the predictors $x_1, \ldots, x_p$ and the outcome $y$. This model can support the reasoning
behind most methods of the dimension reduction type, such as principal components regression (Jolliffe, 2002), partial least squares (Wold et al., 1984) and more recent methods by Burnham et al. (1999a,b) and Bair et al. (2004).

The joint model we propose is a generalization of factor analysis, a latent variable model often used in psychometry (Bartholomew and Knott, 1999). In this model a set of unobserved latent variables \( f_1, \ldots, f_m \) linearly determines both the predictors \( x \) and the outcome \( y \), although both are also subject to error. We assume that the error in \( x \) is independent from the error in \( y \), so that \( y \) is conditionally independent of \( x \) given \( f = (f_1, \ldots, f_m)' \). Graphically:

\[
\begin{array}{c}
\xymatrix{
x_1, \ldots, x_p \\
\ar@{-}[d] \quad \ar@{.}[d] \\
& y}
\end{array}
\quad \begin{array}{c}
f_1, \ldots, f_m \\
\ar^y \ar_{x_1, \ldots, x_p}
\end{array}
\]

We explicitly do not assume that the dimension \( m \) of the latent space is known, but only that it is smaller than the sample size \( n \). Furthermore, our model is more general than the factor analysis model in the sense that we do not assume that the error in \( x_1, \ldots, x_p \) are independent or identically distributed.

In this model we show how to estimate parameters and how to construct a prediction rule for predicting \( y \) from \( x \). We also calculate the mean squared error of prediction for the resulting prediction rule. Based on these calculations, we argue for a prediction rule that first weights the predictors \( x_1, \ldots, x_p \) based on their correlation with \( y \) and then applies a principal components analysis based on this weighting, using only few components. Essentially, this is a generalization of the method proposed by Bair et al. (2004), which was derived on the basis of a very different and much more restrictive model.

To keep the technicalities limited, we have chosen to limit the discussion in this paper to the linear model, in which the outcome \( y \) is continuous and the desired prediction rule is linear. However, the approach is extendable to generalized linear models and we keep the possibility of this extension always in mind.

Before going into the details of the model in section 6.3, in the next section we first investigate some general issues involved in methods of the dimension reduction type and discuss some familiar and less well known methods.

### 6.2 Bias and variance

Dimension reduction methods combat overfit by replacing the original set of predictor variables \( x_1, \ldots, x_p \) with a small set of orthogonal linear combinations of these variables. These linear combinations, sometimes confusingly called ‘latent variables’, are subsequently used to predict the outcome \( y \).
Chapter 6. Model-based dimension reduction

The most basic dimension reduction method is Principal Components Regression (PCR), which uses the first few principal components of the matrix $X$ for prediction (Jolliffe, 2002, Chapter 8). Other important examples of dimension reduction methods include partial least squares (Wold et al., 1984), various types of continuum regression (Abraham and Merola, 2005; Burnham et al., 1996; Stone and Brooks, 1990) and, more recently, methods proposed by Burnham et al. (1999a,b) and by Bair et al. (2004). There are many more examples. All these methods differ from PCR mainly because they also use the training outcomes $y$ to determine the principal components, and they differ from each other in the way they use $y$.

There is a general motivation for all methods of the dimension reduction type. This motivation is best explained through the example of PCR, the mechanics of which are very well known. PCR reduces the matrix $X$ to only its large-variance principal components, ignoring the small-variance principal components for the prediction of $y$. This will always reduce the variance of the prediction, because the regression coefficients of the small-variance principal components are difficult to estimate accurately. However, it may introduce bias, because the small-variance principal components of $X$ might be important predictors of $y$. This is the well-known trade-off between bias and variance (Hwang and Nettleton, 2003).

It follows that PCR has best predictive performance when the bias introduced is small, which happens when the small-variance principal components have little or no predictive value for $y$. A main concern when using PCR is therefore the choice of the number of components: using too many components does not reduce the variance enough, while using too few components may result in missing out those principal components that are important for prediction (Jolliffe, 2002, pp. 173–177). This dilemma is usually solved using methods like cross-validation or AIC, which use the $y$ data to judge which number of components has the best predictive performance. This ‘estimation’ of the number of components reintroduces some variance from $y$ in order to reduce the potential bias.

Other dimension reduction type methods typically introduce the dependence of the choice of the latent variables on $y$ at an earlier stage. They let the latent variables themselves directly depend on the outcome vector $y$, so that the important principal components are more likely to be among the first few latent variables selected. Partial Least Squares (Wold et al., 1984) chooses latent variables that have maximum covariance with $y$ instead of maximum variance. Burnham et al. (1999a,b) choose the latent variables as the eigenvectors corresponding to the largest eigenvalues of the matrix $XX' + \lambda yy'$ for some value of $\lambda$, instead of those of $XX'$ (see also Tan et al., 2005, for an interesting applica-
Chapter 6. Model-based dimension reduction

Bair et al. (2004) propose a pre-selection of the predictor variables based on their correlation with $y$, prior to doing principal components. By using $y$ to choose the latent variables, all these methods essentially reintroduce some variance in order to control the bias.

Ideally, it should follow from a model which method best controls the bias with least reintroduction of variance. Unfortunately, there is little theoretical guidance as to which dimension reduction method is optimal in which situation. The most popular methods of PCR and Partial Least Squares are not based on any model, while the more recent methods of Burnham et al. and Bair et al. are based on relatively restrictive ones (Bair et al., 2004; Burnham et al., 1999a,b).

It is to aid the theoretical discussion on the question which dimension reduction method to use that we formulate our model in section 6.3. This model is the most general model that can motivate a dimension reduction approach. Basically, the only thing it assumes is that a set of latent variables truly exists.

### 6.3 A basic joint model

The basis of our joint model is the graphical model (6.1), which states that a set of latent variables $f_1, \ldots, f_m$ exists, such that $y$ is conditionally independent of $x_1, \ldots, x_p$ given $f_1, \ldots, f_m$. This assumption implies that prediction of $y$ from $x = (x_1, \ldots, x_p)'$ should proceed via ‘prediction’ of the vector $f = (f_1, \ldots, f_m)'$ of latent variables. This property makes it the basic model underlying methods of the dimension reduction type.

To keep the model simple, we only make a few simple extra assumptions. Firstly, like most dimension reduction methods we assume linearity.

$$
\begin{align*}
    y &= \mu_y + \beta' f + \epsilon \\
    x &= \mu + A' f + e
\end{align*}
$$

Here, $\mu_y$ and $\mu$ (a $p$-vector) are the marginal means of $y$ and $x$, respectively. The parameters $\beta$ and $A$ are an $m$-vector and an $m \times p$ matrix of loadings, which determine the relationship between the observed and latent variables. For the error terms $e$ (a $p$-vector) and $\epsilon$ we assume zero mean and variance $E(ee') = \Psi$ and $E(\epsilon^2) = \sigma^2$. By (6.1), $e$ and $\epsilon$ are uncorrelated. Further, for deriving maximum likelihood results we shall assume that the distribution of the errors $e$ and $\epsilon$ is normal. This normality assumption is convenient, but not strictly necessary. The results of this paper can also be rephrased in terms of best linear unbiased prediction (BLUP) under mild conditions. As to the distribution of $f$, we only assume that it has finite mean and covariance matrix, which can then be taken as $0$ and $I$ without loss of generality.
It should be remarked that the model, as presented here, is slightly overparametrized if \( m > 1 \). If \( W \) is an \( m \times m \) orthogonal matrix, replacing \( A \) with \( WA \) and \( \beta \) with \( W\beta \) results in the same joint distribution of \( x \) and \( y \). This means that single parameter values or estimates of \( A \) and \( \beta \) cannot immediately be interpreted, but only functions of \( A \) and \( \beta \) that are invariant to multiplication by \( W \). As our prime purpose is prediction, there is no real need to resolve this overparametrization (see Bartholomew and Knott, 1999, for various methods to choose a rotation \( W \)).

The terms \( \mu_y \) and \( \mu \) can easily be extended to linear regression functions to incorporate predictors that do not fit in the latent variable structure. The estimation of \( \mu_y \) and \( \mu \) (or their regression extensions) and their use in prediction is straightforward, however, and only complicates notation. For simplicity, we shall therefore assume in the rest of the paper that both \( \mu_y \) and \( \mu \) are known. They can then be taken as zero without loss of generality.

The model presented here is very similar to the factor analysis model frequently used in psychometry. The main difference in the model formulation is that we do not require the matrix \( \Psi \) to be diagonal. In psychometry the model is exclusively used in the \( p < n \) situation (Bartholomew and Knott, 1999; Magnus and Neudecker, 1999).

We have formulated the above model in terms of the joint distribution of the predictors \( x \) and the outcome \( y \), because this is much more flexible than a model in terms of the conditional distribution of \( y \) given \( x \), such as a regression model. One aspect of this flexibility is that the fitted joint model can also be used for prediction when there are missing data in \( x_{\text{new}} \). Another flexible aspect of joint modelling is that it is easier to incorporate theoretical knowledge about relationships between variables into a joint model than in a conditional model. This can already be seen in the assumption (6.1) about the existence of latent variables, which can be directly translated into statements about the joint distribution, but not in statements about the conditional distribution. One may also, for example, have knowledge that certain predictors are uncorrelated with the outcome \( y \). This can be immediately incorporated in the joint model as the statement that the corresponding entries of \( A'\beta \) are zero. Such a statement cannot be similarly incorporated into a regression model, because the regression coefficient of a predictor that is uncorrelated with \( y \) does not have to be zero. In general, extending the simple joint model with detailed knowledge about dependency relationships of predictors can be easily done using theory on graphical models. In many cases, however, there is only the vague and implicit, but nevertheless important, assumption that the model has a ‘simple’ structure, which can be translated as the statement that the matrix \( A'A \) and the vector \( A'\beta \) have many zero elements. We come back to this vague model
structure assumption in Section 6.8.

We shall derive most of the results in this paper using “reverse asymptotics”, in which the number of predictors $p$ goes to infinity, while the number of samples $n$ remains fixed or goes to infinity at a much slower rate. We need a few additional assumptions to make these reverse asymptotics well-defined. Essentially, we shall let $p$ grow by simply adding new predictor variables to the vector $x$. This means that as the parameter space grows, the dimensions of the matrices $A$ and $Ψ$ grow. We impose the following two restrictions:

1. There are constants $0 < k \leq K < \infty$ such that all eigenvalues of $Ψ$ are between $k$ and $K$ for all $p$.
2. The limit $\lim_{p \to \infty} \frac{1}{p} AA'$ exists and is of full rank $m$.

First note that by assumption 2 the number of latent variables $m$ does not grow with $p$. The value of $m$ is therefore assumed to be small relative to $p$. For simplicity of notation, we also assume that $m < n$ throughout this paper, but this is not a very critical assumption.

The usefulness of the two assumptions 1 and 2 is that they neatly separate the covariance matrix $A'A + Ψ$ of $x$ into structural covariance ($A'A$) and local covariance ($Ψ$). The structural covariance is caused by a limited number of latent variables, but each affects a number of the predictors that grows with $p$. The local covariance is caused by a vector of errors that grows with $p$, but each independent error term affects only a limited number of predictors.

### 6.4 Regression

From the model equations (6.2) we can easily derive the joint distribution of the observable variables $x$ and $y$. From this joint distribution we can derive any conditional distribution we like. The conditional distribution that is most interesting for prediction is the conditional distribution of $y$ given the whole vector $x$.

The joint vector $z = (y, x')'$ has mean zero and covariance matrix

$$
Σ_z = \left( \begin{array}{cc}
β'β + σ^2 & β'A \\
A'β & A'A + Ψ
\end{array} \right).
$$

The distribution of $z$ is normal if the distributions of $x$ and $y$ are. Therefore, under normality assumptions it follows that $y$ given $x$ is again normal with a mean that is linear in $x$:

$$
E[y \mid x] = γ'x \quad (6.3)
$$
where \( \gamma = (A'A + \Psi)^{-1}A'\beta \) is a vector of regression coefficients. If normality is not assumed, the equation (6.3) gives the best linear unbiased prediction (BLUP) of \( y \) given \( x \).

Using a singular value decomposition on \( A\Psi^{-1/2} \) we can also write the regression coefficients \( \gamma \) as

\[
\gamma = \Psi^{-1}A'(A\Psi^{-1}A' + I)^{-1}\beta.
\] (6.4)

This expression will turn out to be more useful. It is computationally easier, as it does not involve inversion of the complicated \( p \times p \) matrix \( A'A + \Psi \).

### 6.5 Easy prediction

We solve the prediction problem in stages. In the previous section we have seen that in the trivial situation that all parameters are known we should simply use equation (6.3) to predict \( y_{\text{new}} \) from \( x_{\text{new}} \). In this section we study the still relatively easy situation in which the structural parameters \( A \) and \( \beta \) are known (and hence \( m \) is known), but only the error covariance \( \Psi \) is not known.

The great difficulty in this situation is that it is almost impossible to estimate \( \Psi \) accurately enough from the training data. This can already be inferred from the fact that estimating \( \Psi \) means estimating \( p^2 \) parameters, while only \( n(p + 1) \) degrees of freedom are available in the training set. All commonly used estimates of a covariance matrix therefore result in extremely ill-conditioned estimates of \( \Psi \). This ill-conditionedness causes great problems because prediction involves the inverse of the matrix \( \Psi \).

Ledoit and Wolf (2004) studied the general problem of estimating covariance matrices with high-dimensional data. They proposed an estimate that is a linear combination of the naive maximum likelihood estimate and a chosen matrix \( \Theta \) (typically the identity matrix). This biases the estimated covariance matrix towards the chosen matrix \( \Theta \). It also shrinks the eigenvalues of the estimate towards each other and forces them all to be positive, so that the resulting estimate is always invertible. However, as the dimension \( p \) of the covariance grows relative to the sample size \( n \), the bias becomes dominant in the estimate of Ledoit and Wolf. In the limit \( p \to \infty \), for fixed \( n \), the optimal estimate is all bias and no variance. This essentially means that it is hopeless to try to estimate the covariance matrix \( \Psi \) in the \( p \to \infty \), \( n \) fixed situation that we are interested in.

It does not mean, however, that prediction of \( y_{\text{new}} \) from \( x_{\text{new}} \) is hopeless. We can simply use the limiting estimate of Ledoit and Wolf, which is all bias and no variance and take any fixed matrix \( \Theta \) as an ‘estimate’ for \( \Psi^{-1} \). Such a
\( \Theta \) is not truly an estimate so we shall refer to it as a *surrogate*. It should have similar properties for \( p \to \infty \) as \( \Psi^{-1} \). The properties we need are

1. There are constants \( 0 \leq l \leq L < \infty \) such that all eigenvalues of \( \Theta \) are between \( l \) and \( L \) for all \( p \).

2. The limit
   \[
   G = \lim_{p \to \infty} \frac{1}{p} A \Theta A'
   \]
   exists and \( \left( \frac{1}{p} A \Theta A' - G \right)^2 = O(p^{-1}) \).

3. The limit
   \[
   \tau^2 = \lim_{p \to \infty} \frac{1}{p} \text{trace}(\Psi \Theta)
   \]
   exist and \( \left( \frac{1}{p} \text{trace}(\Psi \Theta) - \tau^2 \right)^2 = O(p^{-1}) \).

Note that we allow \( l = 0 \) in property 1, which is importantly different from assumption 1: unlike \( \Psi^{-1} \), the matrix \( \Theta \) is allowed to be singular. Properties 2 and 3 only serve to rule out some very exotic or degenerate choices of \( \Theta \).

Plugging in the surrogate \( \Theta \) for \( \Psi^{-1} \) in the prediction rule (6.3) results in a prediction rule with regression coefficients \( \Theta A'(A \Theta A' + I)^{-1} \beta \) instead of \( \gamma \). As the identity matrix \( I \) is negligible next to \( A \Theta A' \) if \( p \) is large, we can replace this by the simpler expression

\[
\gamma_{\Theta} = \Theta A'(A \Theta A')^{-1} \beta. \tag{6.5}
\]

This vector of regression coefficients \( \gamma_{\Theta} \) will usually be drastically different from the vector \( \gamma \) of ‘true’ regression coefficients. But, surprisingly, the resulting predictions will be precisely the same if \( p \) is large enough. This is stated in Theorem 1.

**Theorem 1** If the matrix \( G \) has full rank \( m \), we have

\[
E[(\gamma'_{\Theta} x - \beta' f)^2 \mid f] = O(p^{-1}).
\]

for every \( f \).

The theorem says that in the limit \( p \to \infty \) using any surrogate \( \Theta \) for \( \Psi^{-1} \) will give perfect predictive performance, as \( \beta' f \) is the expectation of \( y \) given \( f \). We can say that the prediction \( \gamma'_{\Theta} x \) gives a \( p \)-consistent estimate of the mean of \( y \), where \( p \)-consistency is defined completely analogously to ordinary consistency. By the property of the best linear unbiased prediction, using the true \( \Psi^{-1} \) for \( \Theta \) would still give a prediction with least variance for finite \( p \), but the extra
variance caused by using a ‘wrong’ $\Theta$ disappears when $p$ grows to infinity. The difference will therefore be negligible relative to many other sources of estimation and prediction error that we will encounter later.

The easiest way to understand the role of the surrogate $\Theta$ is view it as a weighting matrix. In the setup (6.1) almost every predictor variable in $x$ carries information on $f$ (and through $f$ on $y$). Most of this information is redundant, however, if $p$ is much larger than $m$: a good choice of $m$ predictor variables with small error would be enough to summarize almost all information in $x$ on $y$. If we do use all predictor variables, we are free to choose our own weighting to aggregate their information on $f$. The optimal weighting for finding $f$ from $x$ is $\Psi^{-1}$, by (6.3), which weights each predictor variable inversely to its error variance. However, by Theorem 1 any other weighting which spreads the weight over many predictor variables will do equally well. The intuitive reason for this is the abundance of information on $f$ in $x$ if $p$ is large.

The interpretation of $\Theta$ as a weighting will be useful throughout this paper. Consequently, all estimates and predictors involving $\Theta$ proposed in this paper will be invariant to multiplication of $\Theta$ by a constant. Hence only the relative magnitudes of the entries of $\Theta$ are important, as is appropriate in a weighting matrix.

The fact that different methods may have very different regression coefficients and still result in very similar predictions has been noted before (Burnham et al., 2001). It should be seen as a warning against using regression coefficients as a basis for modelling or interpretation, and another argument for modelling in terms of the joint distribution.

### 6.6 Estimation

There is a big difference between estimation of $\Psi$ and estimation of $A$, the other large matrix of parameters. Estimation of $\Psi$ is difficult, even when $A$ is known, but estimation of $A$ is relatively easy, even if $\Psi$ is unknown. We show this in this section.

Estimation and finding the prediction rule will be based on a training sample: an $n \times p$ matrix $X$ of predictor variables and a corresponding $n$-vector $y$ of outcome variables. Call $F$ the $n \times m$ matrix of the realizations of the unobserved latent variables $f$ for the individuals in the training sample.

We first look at the situation in which both $\Psi$ and $m$ are known. In that situation we can use a standard theorem from factor analysis, formulated by Magnus and Neudecker (1999, Chapter 17, Section 12), which we rephrase here as Theorem 2.
Theorem 2 (Magnus and Neudecker) Let \( \tilde{\Lambda} \) be the \( m \times m \) diagonal matrix of the \( m \) largest eigenvalues of the matrix \( \tilde{S} = \frac{1}{p+1} \{ X \Psi^{-1} X' + \sigma^{-2} y' y \} \) and let \( \tilde{U} \) be the \( n \times m \) orthogonal matrix with the corresponding eigenvectors. If \( \Psi \) and \( m \) are known and the distribution of \( f \) is normal, maximum likelihood estimates of \( A \) and \( \beta \) are given by

\[
\tilde{A} = n^{-1/2} \tilde{Y}^{1/2} \tilde{U}' X \\
\tilde{\beta} = n^{-1/2} \tilde{Y}^{1/2} \tilde{U}' y
\]

where \( \tilde{Y} \) is the \( m \times m \) diagonal matrix with \( \tilde{Y}_{ii} = \max(0, 1 - \frac{n}{p} \tilde{\Lambda}_{ii}^{-1}) \).

Note that the maximum likelihood estimate is not unique if \( m > 1 \), due to the overparametrization mentioned in Section 6.3. If \( W \) is an \( m \times m \) orthogonal matrix, then \( W\tilde{A} \) and \( W\tilde{\beta} \) are also maximum likelihood estimates.

The proof of Theorem 2 is given in Magnus and Neudecker (1999, Chapter 17, Section 12). Their Theorem and its proof are phrased in the context of traditional factor analysis. Consequently, they use the additional implicit assumptions that \( p < n \) and that \( \Psi \) is diagonal. However, it is a simple exercise to show that their proof also holds for \( p \geq n \) and general positive definite \( \Psi \).

The maximum likelihood estimates of Theorem 2 are not immediately usable as they involve the matrix \( \Psi^{-1} \), which cannot be estimated in high-dimensional data. The standard techniques used in factor analysis to estimate \( A, \beta \) and \( \Psi \) simultaneously (Magnus and Neudecker, 1999, Chapter 17, Sections 12–14) cannot therefore be used. However, we can use the same trick that was used in Section 6.5, simply replacing \( \Psi^{-1} \) by a well-conditioned surrogate \( \Theta \).

The estimates \( \tilde{A} \) and \( \tilde{\beta} \) also involve the unknown \( \sigma^2 \). The influence of \( \sigma^2 \) on the estimate disappears very quickly, however, when \( p \) becomes large. In the \( p \to \infty, n \) fixed situation, there is no loss when we simply replace \( \sigma^{-2} \) by zero, just as we replace \( \Psi^{-1} \) by \( \Theta \). This neglect of \( y \) in the estimation of \( A \) stands in sharp contrast the method of Burnham et al. (1999b), which makes \( \sigma^{-2} \) an important tuning parameter. Burnham’s method is very sensible in the applications it was designed for, where both \( x \) and \( y \) are high-dimensional. It is also a sensible strategy when \( p \) is small (Wall and Li, 2003). However, in applications with high-dimensional \( x \) but univariate \( y \), the influence of \( y \) on the estimate \( \tilde{A} \) should always disappear when \( p \to \infty \) and \( \sigma^2 > 0 \).

Aside from the unknown \( \Psi \), there is also the unknown value of \( m \). Just like estimation of \( \Psi \), accurate estimation of the true value of \( m \) is very difficult, if not impossible. The solution we propose to the problem of the unknown \( m \) is similar to the problem of the unknown \( \Psi \). We simply replace the unknown \( m \) with a chosen \( q \geq 0 \). This \( q \) is not trying to estimate the true number of latent
variables; it is the number of latent variables we use for prediction. We shall assume that \( q \leq m \), but this is only to keep notation and proofs simple.

We propose to estimate \( A \) and \( \beta \) as

\[
\hat{A} = n^{-1/2} \hat{U}'X \\
\hat{\beta} = n^{-1/2} \hat{U}'y
\]

(6.6)

Where \( \hat{U} \) and is defined as the \( n \times q \) orthogonal matrix with the eigenvectors corresponding to the \( q \) largest eigenvalues of the matrix

\[
S = \frac{1}{p} X \Theta^r X'.
\]

The motivation for these estimates of \( A \) and \( \beta \) will come from asymptotic arguments very similar to the arguments used in Section 6.5. Theorem 3 shows that when \( p \to \infty \) it makes no difference whether we use the matrix \( \tilde{S} \) involving the true \( \Psi \) and \( \sigma^2 \) or the matrix \( S \) involving the surrogate \( \Theta \).

Because the matrix \( S \) does not involve \( y \), the estimates \( \hat{A} \) and \( \hat{\beta} \) can easily be adapted to the situation where \( y \) does not have a normal distribution, but depends on \( f \) through a generalized linear model. In that case \( \beta \) is estimated as the regression coefficients of the Generalized Linear Model with outcome \( y \) and \( n^{1/2} \hat{U} \) as the design matrix (see also Bair et al., 2004).

By \( \| \cdot \| \) we denote the Frobenius norm: \( \| C \| = \text{trace}(C'C)^{1/2} \).

**Theorem 3** If \( q \leq \text{rank}(G) \), there is an \( m \times q \) semi-orthogonal matrix \( V \), depending on \( \hat{A} \) and \( \tilde{A} \), such that, almost surely in \( F \),

\[
E[\| \hat{\beta} - V'\tilde{\beta} \|^2 | F] = O(p^{-1}) \\
p^{-1}E[\| \hat{A} - V'\tilde{A} \|^2 | F] = O(p^{-1})
\]

If \( q = m \) or if the \( q \)-th and \( q + 1 \)-th eigenvalues of \( G = \lim_{p \to \infty} p^{-1} A \Theta^r A' \) are distinct, the result holds uniformly in \( n \).

Theorem 3 states that if \( p \) is large enough and \( q = m \), the estimates \( \hat{\beta} \) and \( \tilde{\beta} \) and \( \hat{A} \) and \( \tilde{A} \) differ only by a rotation. As a rotation of \( \hat{A} \) is also a maximum likelihood estimate of \( A \), this means that if \( q = m \), the estimates (6.6) are asymptotically equivalent (for \( p \to \infty \)) to the maximum likelihood estimates of \( A \) and \( \beta \). The formulae of (6.6) thus give a good complementary estimation procedure to the traditional iterative estimation procedure for \( A \) and \( \beta \) used in factor analysis models (Magnus and Neudecker, 1999). The procedure is complementary because it works precisely in the high-dimensional situation in which the traditional procedure does not.
When interpreting Theorem 3 one must keep in mind that estimates which differ only by a $q \times q$ rotation matrix can be considered as equivalent because they lead to the same estimated joint distribution. If $q = m$, therefore, estimates $\hat{A}$ and $\hat{\beta}$ for different surrogates $\Theta$ are all asymptotically equivalent because in the limit $p \to \infty$ they only differ by a $q \times q$ rotation matrix $V$. However, if $q < m$, estimates $\hat{A}$ and $\hat{\beta}$ for different surrogates are different even in the limit $p \to \infty$, because then the $m \times q$ matrices $V$ differ from each other by more than a $q \times q$ rotation.

The effect of the surrogate $\Theta$ on the estimate is best understood in terms of weighted principal components. The matrix $\hat{U}$ is the standardized matrix of the first $q$ principal components of the weighted data matrix $X\Theta^{1/2}$. If $p \to \infty$ the principal components of $X\Theta^{1/2}$ will be the same as those of $FA\Theta^{1/2}$ (see the proof of Theorem 3). The combined span of the first $m$ principal components of $FA\Theta^{1/2}$ does not depend on $\Theta$, as it is simply the column span of $F$. However, the principal component variances do depend on $\Theta$. Therefore the span of the $q$ principal components with largest variance does depend on $\Theta$. For finite $p$, the minimum variance estimate still remains the estimate with $\Theta = \Psi^{-1}$ (Wentzell et al., 1997).

We have so far only proved that we can define alternative estimates of $A$ and $\beta$ which are non-iterative and are as good as the maximum likelihood estimates for known $\Psi$ if $p \to \infty$. Of course, this only shows that the proposed estimate is good if the maximum likelihood estimate itself is good, which may not be the case if $m$ is close to $n$. However, in the next section we show that the estimate $\hat{A}$ has good properties when used for prediction.

### 6.7 Prediction

We want to predict the outcome $y$ from the predictors $x$ in the model (6.2) in which all parameters are unknown. We propose to combine the results from the previous sections by plugging the estimates of Section 6.6 into the prediction rule of Section 6.5.

This would lead to the prediction rule predicting $y_{\text{new}}$ with $\hat{\gamma}'x_{\text{new}}$ where the regression coefficients are

$$\hat{\gamma} = \Theta\hat{A}'(\hat{A}\Theta\hat{A}')^{-1}\hat{\beta}.$$ 

Compare (6.5). Note that if we take $\Theta = I$ the regression coefficients $\hat{\gamma}$ are exactly the regression coefficients from a principal components regression. This can be seen by writing $\hat{\gamma}'x_{\text{new}} = y'\hat{U}\hat{U}'(XX')^{-1}\hat{U}\hat{U}'Xx_{\text{new}}$. If we take $\Theta \neq I$, the vector $\hat{\gamma}$ is the vector of regression coefficients of a weighted principal components regression.
This prediction rule needs an adjustment for the large $p$, small $n$ situation that we are interested in. It turns out that when $n$ is small, the prediction $\tilde{\gamma}'x_{\text{new}}$ tends to induce too much shrinkage. This excessive shrinkage is caused by overfit of $\hat{A}$ to the noise in $X$. This overfit causes $\hat{A}\hat{\Theta}\hat{A}'$ to be systematically larger than $A\Theta A'$: we have

$$
\lim_{p \to \infty} \frac{1}{p} \hat{A}\hat{\Theta}\hat{A}' = \frac{1}{p} A\Theta A' + \frac{1}{n} \tau^2 I,$$

where $\tau^2 = \lim_{p \to \infty} p^{-1}\text{trace}(\Theta\Psi)$. We propose to remedy this overfit in the small $n$ situation by subtracting an estimate of $\tau^2 I$. This leads to the following prediction rule:

$$
\hat{y}_{\text{new}} = \hat{\gamma}'x_{\text{new}} \quad (6.7)
$$

with regression coefficients

$$
\hat{\gamma} = \Theta \hat{A}'(\hat{A}\hat{\Theta}\hat{A}' - \frac{p}{n} \tau^2 I)^{-1}\hat{\beta}. \quad (6.8)
$$

The difference between the prediction based on $\hat{\gamma}$ and on $\tilde{\gamma}$ disappears very quickly when $n$ becomes large, but can be important in the small $n$ situation. It is essential for the asymptotic $p \to \infty$ result in Theorem 4.

The adjusted prediction rule (6.7) involves a new quantity $\hat{\tau}^2$, which is an estimate of $\tau^2$. We define $\hat{\tau}^2$ as

$$
\hat{\tau}^2 = \frac{1}{n-r}\text{trace}(S\hat{Q})
$$

where $\hat{Q}$ is the rank $n-r$ projection matrix for projection on the eigenvectors corresponding to the $n-r$ smallest eigenvalues of $S$. It is shown in the appendix that $\hat{\tau}^2 \to \tau^2$ as $p \to \infty$ uniformly in $n$ whenever $r > m$, so that $\hat{\tau}^2$ is a good estimate of $\tau^2$ when $p$ is large, even when $n$ is small. It is easily checked that the matrix $\hat{A}\hat{\Theta}\hat{A}' - \frac{p}{n} \hat{\tau}^2 I$ in (6.8) is always positive (semi-)definite.

The most important property of the prediction $\hat{y}_{\text{new}}$ is Theorem 4.

**Theorem 4** If $p \to \infty$,

$$
\text{E}[ (\hat{y}_{\text{new}} - \bar{y}_{\text{new}})^2 | F] = O(p^{-1})
$$

almost surely, where

$$
\bar{y}_{\text{new}} = y'F(F'F)^{-1/2}VV'(F'F)^{-1/2}f_{\text{new}}
$$

and $V$ is the $m \times q$ semi-orthogonal matrix of the eigenvectors of the $q$ largest eigenvalues of the matrix $(F'F)^{1/2}G(F'F)^{1/2}$. If $q = m$ or if the $q$-th and $q + 1$-th eigenvalues of $G = \lim_{p \to \infty} p^{-1} A\Theta A'$ are distinct, the result holds uniformly in $n$. 

100
Note that if \( q = m \), the \( \tilde{y}_{\text{new}} \) in Theorem 4 is simply the least squares prediction of \( y_{\text{new}} \) in the situation where \( F \) and \( f_{\text{new}} \) are observed variables instead of latent variables. If \( q < m \), the projection matrix \( VV' \) introduces bias into the prediction, because not all latent variables are used for prediction.

The prediction based on \( \tilde{y}_{\text{new}} \) is not perfect: it has bias if \( q < m \), and it also has a prediction variance. The bias and the prediction variance of \( \tilde{y}_{\text{new}} \) do not vanish when \( p \to \infty \). By Theorem 4, therefore, the prediction error of \( \hat{y}_{\text{new}} \) will be dominated by the prediction error of \( \tilde{y}_{\text{new}} \) if \( p \) is large and \( n \) is small, while the difference between \( y_{\text{new}} \) and \( \tilde{y}_{\text{new}} \) will be negligible. We must therefore study the prediction error of our prediction \( \hat{y}_{\text{new}} \) through the prediction error of \( \tilde{y}_{\text{new}} \).

The variance and bias of the prediction \( \tilde{y}_{\text{new}} \) are easy to calculate conditional on \( F \) and \( f_{\text{new}} \). The variance \( v^2 = \text{Var}[\tilde{y}_{\text{new}} \mid F, f_{\text{new}}] \) is
\[
v^2 = \sigma_y^2 \| V'(F'F)^{-1/2} f_{\text{new}} \|^2 \tag{6.9}\]
and the bias \( b = \text{E}[\tilde{y}_{\text{new}} - \beta' f_{\text{new}} \mid F, f_{\text{new}}] \) is
\[
b = \beta' (F'F)^{1/2} (I - VV') (F'F)^{-1/2} f_{\text{new}}. \tag{6.10}\]

These results would be easier to interpret if we could take the expectation of the variance and of the squared bias over \( F \) and \( f_{\text{new}} \). However, there is no analytical solution for these expectations for finite \( n \), mainly because of the complicated dependence between \( V \) and \( F \). However, the small \( n \) behaviour of the bias and variance of \( \tilde{y}_{\text{new}} \) is very similar to the large \( n \) behaviour. As \( n \to \infty \),
\[
E[v^2] \to q \sigma^2 \\
E[b^2] \to \beta' (I - VV') \beta \tag{6.11}\]
where \( V \) is now a matrix of eigenvectors of \( G \).

A trade-off between \( v^2 \) and \( b \) determines the performance for different \( q \) and \( \Theta \) of the prediction rule (6.7) proposed above. Some interesting conclusions can be drawn.

In the first place it is not always optimal to take \( q = m \), even if \( m \) is known. Reducing \( q \) below \( m \) will usually increase the squared bias \( b^2 \), but it will always decrease the prediction variance \( v^2 \). If the decrease in variance is larger than the increase in bias, the prediction rule with smaller \( q \) will be the better one. This shows that it is often not worthwhile to try to estimate \( m \), as knowledge of \( m \) does not necessarily lead to improved prediction accuracy.

There is a value of \( q \) somewhere between 0 and \( m \) where the trade-off between bias and variance leads to optimal prediction error. The location of this optimum depends on \( n \): larger \( n \) means smaller prediction variance \( v^2 \), but not
a smaller bias \( b \). Therefore the optimal trade-off will be different. Typically, a larger \( n \) will lead to a larger optimal \( q \). The location of the optimal \( q \) also depends on \( \Theta \), as the choice of \( \Theta \) also affects the size of the bias.

Unlike the choice of \( q \), the choice of \( \Theta \) does not involve a trade-off between bias and variance. The reason for this is that, as remarked above, the choice of \( \Theta \) only influences the distribution of the bias \( b \) of the prediction, but not of the variance \( \nu^2 \). Therefore, we can choose a \( \Theta \) that makes the bias small, without automatically incurring a large prediction variance. Furthermore, if we can find a \( \Theta \) that produces small bias even for small values of \( q \), we can choose \( q \) small, thereby indirectly reducing the variance \( \nu^2 \).

The prediction rule (6.7) therefore has a predictive performance that is as good, if \( p \to \infty \), as the prediction \( \hat{y}_{\text{new}} \) of an ‘oracle’ that observes the unobservable latent variables. We get this predictive performance for any \( \Theta \) when we choose \( q = m \). A smart choice of \( \Theta \) and \( q \) may even result in a better predictive performance than \( \hat{y}_{\text{new}} \). However, whereas finding the optimal \( q \) for fixed \( \Theta \) is doable, finding an optimal \( \Theta \) is a daunting task even for fixed \( q \), due to the enormously large search space. We discuss one promising strategy in the next section.

### 6.8 Supervised Principal Components

The model of this paper seemed to be especially designed for supporting Principal Components Regression. In the previous section, however, we have already shown that even when \( m \) is known, Principal Components Regression with \( m \) components is not necessarily the optimal prediction rule. In this section we show that there are good arguments in favour of a data-driven way to choose \( \Theta \), which leads to a variant of the Supervised Principal Components method recently proposed by Bair et al. (2004). Due to the increased complexity of a having a random \( \Theta \), exact statements are difficult to prove and this section will be more informal than the previous sections.

Which choice of \( \Theta \) induces the smallest bias? This is easiest to see in the large \( n \) situation of equation (6.11), but it holds similarly for the more complex situation of equation (6.10). The bias is small if \( \beta' (I - V V') \beta \) is small, which happens if \( \beta \) is in the span of the eigenvectors of the \( q \) largest eigenvalues of \( G = \lim_{p \to \infty} p^{-1} A \Theta A' \). If \( \Theta \) is diagonal, with \( i \)-th diagonal element \( \theta_i \), we have

\[
G = \frac{1}{p} \sum_{i=1}^{p} \theta_i \alpha_i \alpha_i'
\]

where \( \alpha_i \) is the \( i \)-th column of \( A \) (an \( m \)-vector). To push the eigenvectors of the large eigenvalues of \( G \) in the direction of \( \beta \), we should give a large weight \( \theta_i \) to
predictor variables which have a large correlation between $\alpha_i$ and $\beta$, and vice versa give a small weight $\theta_i$ to variables which have a low correlation.

As $\Lambda$ and $\beta$ are not known, a suitable proxy to the correlation between $\alpha_i$ and $\beta$ is the correlation between $x_i$ and $y$, where $x_i$ is the $i$-th column of $X$. Predictor variables which have a large correlation between $x_i$ and $y$ tend to have a large correlation between $\alpha_i$ and $\beta$, combined with a small error variance.

This suggests a very simple method for choosing $\Theta$, which can be expected to lead to a small bias $b$ even when $q$ is small. This method finds a limited number of predictors which have the largest correlation with the outcome. It gives these predictors equal weight ($\theta_i = 1$), and all other predictors zero weight ($\theta_i = 0$). Using this $\Theta$ in combination with a small value of $q$ in the prediction rule (6.7) gives precisely the Supervised Principal Components method proposed by Bair et al. (2004), except for the improvement based on $\hat{\tau}^2$ that we proposed in equation (6.7). Variants of Supervised Principal Components can easily be thought of, for example taking $\theta_i$ equal to the squared correlation between $x_i$ and $y$. But these variants are not essentially different, so we study the original proposal by Bair et al. (2004).

For such a “supervised” data-driven choice of a surrogate the discussion on variance and bias in Section 6.7 is not strictly valid, because due to the data-dependent construction of $\Theta$ the matrix $V$ is not independent of the error in the outcomes $y$. Therefore the distinction between bias and variance is not the same as it is for fixed $\Theta$. The data-dependent choice of $\Theta$ will, therefore, usually not just reduce the bias but also increase the variance of the prediction. There is no explicit expression for bias and variance in this case.

Similarly, Theorem 4 does not hold for a $\Theta$ that depends on $X$. However, we can conclude from Theorem 4 that the prediction result is extremely robust against the choice of $\Theta$, because that theorem holds for every fixed $\Theta$. If $\Theta$ spreads the weight over many predictor variables, the first few principal components are principal components of a large subset of the columns of the matrix $X$. By Theorem 4, prediction based on the principal components of any such large subset is indistinguishable from prediction based on the true latent variables $f$. Because this result is result holds for all fixed $\Theta$, we can also expect it to hold for a data-dependent $\Theta$, as long as $\Theta$ is not ‘too data-dependent’, but still selects a large subset of the predictors. We can therefore expect good prediction results if we combine a data-dependent $\Theta$ as in Supervised Principal Components regression, provided $\Theta$ does not put all its weight on a small subset of the predictors.

Using a data-dependent weighting $\Theta$ becomes even more attractive if $n$ is large, because the variance of $\Theta$ will disappear as $n$ grows. For large $n$ the approximations (6.11) of bias and variance are asymptotically still valid. For
large $n$, the Supervised Principal Components choice of $\Theta$ combined with a small value of $q$ will lead to a small prediction variance and a small bias.

Supervised principal components uses the information in $y$ to choose a weighting of the predictor variables to be used to calculate principal components, but does not let $y$ influence these components directly. Therefore it has the desirable tendency to choose those principal components that have a relationship with $y$. However it is relatively robust against pushing the principal components into the directions of the error of $y$, because only a small number of predictors will have their errors correlated with the error of $y$. These will typically be too few in number to have a large influence on the first few principal components.

In many applications, for example in microarray data, there is the implicit assumption that many of the predictor variables are not correlated with the outcome $y$. This can be translated as the assumption that many entries of $A^T \beta$ are exactly zero. Predictors with $\alpha'_i \beta$ zero should ideally be given zero weight when doing the principal components calculations, because they only add weight to unimportant principal component directions. A good way to filter out the predictors with $\alpha'_i \beta = 0$ is to remove the predictors with low correlation with the outcome. This assumption that the model is structured in the sense that many predictor variables are uncorrelated with the outcome therefore also gives an argument in favour of using Supervised Principal Components.

### 6.9 Application

In this section we present a simulation study to illustrate the findings above. To make the simulations realistic, they are based on a microarray gene expression data set of breast cancer patients by Van de Vijver et al. (2002).

Our simulations complement the extensive simulation and data analysis presented by Bair et al. (2004). In their analysis they compare Supervised Principal Components to several other commonly used methods, showing that Supervised Principal Components has good performance relative to other methods. In these simulations they always chose $q = 1$, but chose the number of selected genes using cross-validation.

Our simulation setup is as follows. We used the data by Van de Vijver et al. (2002) in order to obtain realistic values for the parameters $A$ and $\Psi$. In the original data there were 295 patients and 24,481 predictors, of which 293 patients and 23,862 predictors remained after removing some of the patients and predictors due to missing values. We took 10 as the true value of $m$ and estimated $A$ using the procedures described in Section 6.6. To obtain more highly structured relationships between the latent variables $f$ and the observed $x$, as
is biologically expected for microarray data, we applied hard thresholding to $A$, setting all but the 10% largest absolute values of $\hat{A}$ to zero. We took $\Psi$ as a diagonal matrix, the standard method-of-moments estimate of a diagonal $\Psi$ given $A$: 

$$\hat{\Psi} = \frac{1}{n-m} \text{diag}(X'X - n\hat{A}'\hat{A}).$$

These values of $A$ and $\Psi$ were subsequently used as the true values in the simulation experiments that follow.

In our simulation experiment we compared the performance of our version of the Supervised Principal Components method for all values of $q$ and for various choices of the number $s$ of selected predictors. We investigated what values of $q$ and $s$ tend to do well in prediction. We also investigated how this answer depends on $\beta$. We generated datasets based on different choices of $\beta$, each letting $y$ depend on a different latent variable. In the simulation $k$, we define $\beta_k = e_k$, the $k$-th standard basis vector. Depending on the choice of $k$, between 5,742 ($k = 2$) and 989 ($k = 10$) predictors were correlated with the outcome $y$. We always chose $\sigma^2 = 1$, so that at most 50% of the variance of $y$ can be predicted from $x$.

Based on the $A$ and $\Psi$ values chosen above, we generated a training data set of $X$ and $y$ of $n = 20$ patients and $p = 23,862$ predictors using a matrix $F$ of latent variables and based on the model equations (6.2). The performance of the prediction rules created on this data-set was evaluated using an independently generated test set of $n = 100$ patients. The performance of the method depends on the realized values of the latent variables $F$ in the training and test set. Therefore we repeated the whole procedure of generating training and test sets 100 times to be able to average out the effects of $F$. The construction of $\Theta$ for the pre-selection of genes was done for each data set separately.

In the tables we give the results of the simulations for the five different choices of $\beta$. It has to be remarked that due to the thresholding the rows of $A$ are not orthogonal anymore and that the ordering of the norms of the rows may have changed. However, the value of $k$ still gives a good indication how much of the variance of $x$ is explained by the latent variable $f'\beta$: the larger $k$, the more variance $f'\beta$ explains.

The results of the simulations are given in Tables 1–5. From these simulations we see clearly that it is highly worthwhile to make a pre-selection of predictors prior to doing Principal Components Regression. The lowest prediction error of the methods that select fewer than 23,862 predictors is in all cases lower than the prediction error of the ‘plain’ Principal Components Regression. Furthermore, this optimum is attained at much lower values of $q$. However, taking a too small subset to do the Principal Components Regression leads to
Table 6.1: Mean squared prediction error for Supervised Principal Components for various choices of the number of components used and of the number of predictors selected. Simulated data in which the first latent variable is related to y and to 4,201 predictor variables.

<table>
<thead>
<tr>
<th># pred.</th>
<th># components q</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>23,862</td>
<td></td>
<td>0.44</td>
<td>0.28</td>
<td>0.24</td>
<td>0.23</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>10,000</td>
<td></td>
<td>0.18</td>
<td>0.17</td>
<td>0.18</td>
<td>0.19</td>
<td>0.21</td>
<td>0.22</td>
<td>0.23</td>
<td>0.24</td>
<td>0.24</td>
<td>0.25</td>
</tr>
<tr>
<td>5,000</td>
<td>0.14</td>
<td>0.16</td>
<td>0.18</td>
<td>0.20</td>
<td>0.22</td>
<td>0.23</td>
<td>0.23</td>
<td>0.24</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>1,000</td>
<td>0.12</td>
<td>0.14</td>
<td>0.16</td>
<td>0.18</td>
<td>0.20</td>
<td>0.20</td>
<td>0.21</td>
<td>0.21</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
</tr>
<tr>
<td>200</td>
<td>0.13</td>
<td>0.14</td>
<td>0.16</td>
<td>0.17</td>
<td>0.18</td>
<td>0.20</td>
<td>0.21</td>
<td>0.22</td>
<td>0.22</td>
<td>0.22</td>
<td>0.24</td>
</tr>
<tr>
<td>50</td>
<td>0.18</td>
<td>0.19</td>
<td>0.20</td>
<td>0.22</td>
<td>0.23</td>
<td>0.25</td>
<td>0.27</td>
<td>0.28</td>
<td>0.30</td>
<td>0.32</td>
<td>0.32</td>
</tr>
</tbody>
</table>

Table 6.2: Mean squared prediction error for Supervised Principal Components for various choices of the number of components used and of the number of predictors selected. Simulated data in which the second latent variable is related to y and to 5,742 predictor variables.

<table>
<thead>
<tr>
<th># pred.</th>
<th># components q</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>23,862</td>
<td></td>
<td>0.80</td>
<td>0.43</td>
<td>0.35</td>
<td>0.33</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
<td>0.31</td>
</tr>
<tr>
<td>10,000</td>
<td></td>
<td>0.31</td>
<td>0.25</td>
<td>0.26</td>
<td>0.28</td>
<td>0.29</td>
<td>0.30</td>
<td>0.31</td>
<td>0.32</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>5,000</td>
<td>0.22</td>
<td>0.23</td>
<td>0.26</td>
<td>0.28</td>
<td>0.29</td>
<td>0.30</td>
<td>0.31</td>
<td>0.31</td>
<td>0.32</td>
<td>0.32</td>
<td>0.32</td>
</tr>
<tr>
<td>1,000</td>
<td>0.14</td>
<td>0.22</td>
<td>0.24</td>
<td>0.26</td>
<td>0.26</td>
<td>0.27</td>
<td>0.27</td>
<td>0.28</td>
<td>0.28</td>
<td>0.28</td>
<td>0.29</td>
</tr>
<tr>
<td>200</td>
<td>0.14</td>
<td>0.19</td>
<td>0.21</td>
<td>0.23</td>
<td>0.24</td>
<td>0.25</td>
<td>0.26</td>
<td>0.27</td>
<td>0.28</td>
<td>0.28</td>
<td>0.28</td>
</tr>
<tr>
<td>50</td>
<td>0.19</td>
<td>0.22</td>
<td>0.23</td>
<td>0.24</td>
<td>0.25</td>
<td>0.27</td>
<td>0.28</td>
<td>0.30</td>
<td>0.31</td>
<td>0.33</td>
<td>0.33</td>
</tr>
</tbody>
</table>

Table 6.3: Mean squared prediction error for Supervised Principal Components for various choices of the number of components used and of the number of predictors selected. Simulated data in which the third latent variable is related to y and to 2,718 predictor variables.

<table>
<thead>
<tr>
<th># pred.</th>
<th># components q</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>23,862</td>
<td></td>
<td>0.98</td>
<td>0.86</td>
<td>0.61</td>
<td>0.51</td>
<td>0.47</td>
<td>0.46</td>
<td>0.45</td>
<td>0.44</td>
<td>0.44</td>
<td>0.43</td>
</tr>
<tr>
<td>10,000</td>
<td></td>
<td>0.60</td>
<td>0.46</td>
<td>0.38</td>
<td>0.39</td>
<td>0.40</td>
<td>0.41</td>
<td>0.42</td>
<td>0.43</td>
<td>0.43</td>
<td>0.44</td>
</tr>
<tr>
<td>5,000</td>
<td>0.45</td>
<td>0.39</td>
<td>0.36</td>
<td>0.38</td>
<td>0.40</td>
<td>0.41</td>
<td>0.42</td>
<td>0.43</td>
<td>0.43</td>
<td>0.43</td>
<td>0.43</td>
</tr>
<tr>
<td>1,000</td>
<td>0.33</td>
<td>0.32</td>
<td>0.33</td>
<td>0.35</td>
<td>0.38</td>
<td>0.38</td>
<td>0.39</td>
<td>0.40</td>
<td>0.40</td>
<td>0.40</td>
<td>0.41</td>
</tr>
<tr>
<td>200</td>
<td>0.31</td>
<td>0.31</td>
<td>0.33</td>
<td>0.34</td>
<td>0.35</td>
<td>0.36</td>
<td>0.37</td>
<td>0.38</td>
<td>0.39</td>
<td>0.40</td>
<td>0.40</td>
</tr>
<tr>
<td>50</td>
<td>0.37</td>
<td>0.35</td>
<td>0.36</td>
<td>0.37</td>
<td>0.38</td>
<td>0.40</td>
<td>0.40</td>
<td>0.43</td>
<td>0.44</td>
<td>0.44</td>
<td>0.46</td>
</tr>
</tbody>
</table>

inflated prediction error, because the selected data set cannot be called high-dimensional anymore. This is all precisely as expected from the discussion in
Section 6.8. It is interesting to note, however, that even when a pre-selection of genes gives better results, it is not always at very low values of \( q \). This can be seen in Table 6.5. The differences between methods are small here, however.

A first recommendation from these simulations is that a supervised version of Principal Components Regression is always preferable to a non-supervised one, because the optimal prediction error was never attained without pre-selection of genes. It is already common practice in microarray research to do a pre-selection of genes based on the variance of genes. A pre-selection that also takes correlation with the outcome into account may increase predictive accuracy. A second recommendation is that, when using Supervised Principal Components, it is not advisable to only look at the case \( q = 1 \) as Bair et al. do. Often a better prediction error can be found at slightly larger values of \( q \), so it can be worthwhile to also search for an optimum among these values. Cross-validation can be a good strategy for that.
6.10 Discussion

We have constructed a basic joint model of the predictor variables \( x \) and the outcome \( y \), in which both \( x \) and \( y \) depend on a set of \( m \) latent variables. This model is very general because it assumes a general error structure for \( x \) and it does not assume that number of latent variables is known.

We have shown that assuming this model and constructing a prediction rule which has good properties in the \( p \to \infty, n \) fixed, situation leads to weighted principal components regression in a natural way. The ideal weighting is one that puts most weight on those predictor variables that are correlated with the outcome \( y \). This gives good arguments for using a variant of the method of Supervised Principal Components (Bair et al., 2004), which puts all weight on a subset of the predictor variables that is correlated with the outcome.

This result may be considered surprising, because the method of Supervised Principal Components was originally motivated using a very different and much more restrictive model and its good properties were proved in this model using traditional \( n \to \infty \) asymptotics. Furthermore, the model of this paper is actually very similar to the method presented by Burnham et al. (1999a,b) to motivate their method.

The essential assumption for the construction of the method in this paper are that the number of predictors \( p \) is very large, specifically much larger than the sample size \( n \), and that the unpredictable part of \( y \) is not negligible. These assumptions are very realistic in many statistical applications in modern science, where extremely high-dimensional data are become the rule rather than the exception.

The model presented in this paper also makes a few other assumptions which are not strictly essential, but merely serve to keep the subject matter from becoming too technical. Examples of such assumptions are the normality of the errors, the assumptions that \( m < n \) and that \( q \leq m \). We expect that these assumptions can be dropped without leading to important difficulties. Of greater practical relevance is to investigate what happens when the surrogate \( \Theta \) used for estimation of \( A \) and \( \beta \) is different from the surrogate that is used for prediction. This allows more flexibility in the prediction process, for example when there are missing data in \( x_{\text{new}} \). Another important extension of the model would allow \( y \) to depend on \( f \) through a generalized linear model. As Supervised Principal Components is also advocated for these situations, it is interesting to check whether the arguments presented in this paper still apply. At what cost all the above assumptions can be dropped may be an interesting subject for further research.
6.11 Proofs of the theorems

Theorem 1

Proof: Recall that \( x = A'f + e \), so

\[
\gamma'x = \beta'(A\Theta A')^{-1}A\Theta A'f + \gamma'e
\]

so the mean of \( \gamma'x \) is \( \beta'f \). The variance is

\[
\text{Var}(\gamma'x) = \beta'(A\Theta A' + I)^{-1}A\Theta \Psi A'(A\Theta A' + I)^{-1}\beta.
\]

This latter expression is \( O(p^{-1}) \) by Assumptions 1 and 2 and the properties of the surrogate \( \Theta \).

Theorem 3

The proofs of Theorem 3 require the following three Lemmas, which relate the matrix \( S \) to its limiting expectation

\[
\Sigma = F G_{\Theta} F' + \tau^2 I
\]

where \( G_{\Theta} = \lim_{p \to \infty} p^{-1} A\Theta A' \) and \( \tau^2 = \lim_{p \to \infty} p^{-1} \text{trace}(\Theta \Psi) \). The matrix \( \Sigma \) exists and is finite by the properties of \( \Theta \).

Lemma 1 As \( p \to \infty \),

\[
\frac{1}{n^2} E[\|S - \Sigma\|^2 | F] = O(p^{-1}).
\]

The statement holds almost surely in \( F \) uniformly in \( n \).

This lemma is modification of Lemma 1 in Van Houwelingen and Schipper (1981).

Proof: We first prove this lemma under slightly different assumptions. We shall assume that \( \Theta = I \) and \( \Psi \) is diagonal, but not necessarily invertible. Let \( \psi_i^2 \geq 0 \) be the \( i \)-th diagonal element of \( \Psi \), \( x_i \) the \( i \)-th column of \( X \) (an \( n \)-vector) and \( \alpha_i \) the \( i \)-th column of \( A \) (an \( m \)-vector).

Call \( R = S - E(S | F) \). Then we have

\[
R = \frac{1}{p} \sum_{i=1}^{p} (x_i x_i' - F \alpha_i \alpha_i' F' - \sigma_i^2 I).
\]
Given $F$, $R$ is an average of independent zero mean random variables, so by the law of large numbers and the fact that $\text{E}(S \mid F) = \Sigma + O(p^{-1/2})$, by Assumption 2 on page 95, the statement of the lemma follows for fixed $n$.

To prove uniformity in $n$ we will show that

$$\frac{1}{n^2} \text{E}[\|R\|^2 \mid F] = \frac{1}{n^2} p^2 \sum_{i=1}^{p} \text{trace}\{\text{E}[R_i^2 \mid F]\}$$

is bounded in $n$, where $R_i = x_i x_i' - F \alpha_i \alpha_i' F' - \sigma_i^2 I$. For each $i$ we can write $x_i = F \alpha_i + e_i$ with $e_i$ an $n$-vector of independent normal errors with fourth moment $\kappa_i < \infty$. Then

$$\frac{1}{n^2} \text{trace}\{\text{E}[R_i^2 \mid F]\} = \frac{1}{n^2} \text{trace}\{(F \alpha_i e_i + e_i \alpha_i' F' + e_i e_i' - \sigma_i^2 I)^2\}$$

$$= \frac{4}{n} \sigma_i^2 \alpha_i' F \alpha_i + 2 \sigma_i^4.$$

For every $i$ this expression converges a.s. to a limit as $n \to \infty$, because by the strong law of large numbers $n^{-1/2} F F' \to I$ (a.s.). This proves the lemma for $\Theta = I$ and $\Psi$ diagonal.

For general $\Theta$ and $\Psi$, we write $\Psi^{1/2} \Theta^{1/2}$ in a singular value decomposition as $\Psi^{1/2} \Theta^{1/2} = Q D^{1/2} T'$, where $Q$ and $T$ are $p \times p$ orthogonal matrices and $D$ is diagonal. Transform the data matrix $X$ to $\tilde{X} = X \Theta^{1/2}$. Then $\tilde{X}$ and $y$ conform to the general model (6.2) with parameters $\tilde{A} = A \Theta^{1/2} T$ and $\tilde{\Psi} = D$. For this model we can take $\tilde{\Theta} = I$ and apply this lemma on $\tilde{S} = p^{-1} \tilde{X} \tilde{\Theta} \tilde{X}'$ and its corresponding $\tilde{\Sigma}$. This immediately proves the statement of the lemma for $X$, as $\tilde{S} = S$ and $\tilde{\Sigma} = \Sigma$. \hfill $\square$

For the next lemma, define $\hat{P} = \hat{U} \hat{U}'$ and, analogously, define $P$ as the projection matrix for projection on the space spanned by the eigenvectors of the $q$ largest eigenvalues of $\Sigma$. Let $U$ be an $n \times q$ semi-orthogonal matrix such that $P = U U'$.

**Lemma 2** If $q \leq \text{rank}(G)$, the matrix $P$ exists a.s. and is given by

$$P = F (F' F)^{-1/2} V V' (F' F)^{-1/2} F',$$

where $V$ is the $m \times q$ semi-orthogonal matrix of the eigenvectors of $q$ largest eigenvalues of $(F' F)^{1/2} G (F' F)^{1/2}$.

**Proof:** Note that $T = F (F' F)^{-1/2}$ is a.s. an $n \times m$ semi-orthogonal matrix and that $\Sigma$ a.s. has distinct $q$-th and $q + 1$-th eigenvalues. Decompose

$$\Sigma = T (F' F)^{1/2} G (F' F)^{1/2} T' + \tau^2 I.$$
Diagonalize \((F'F)^{1/2}G(F'F)^{1/2} = VDV'\). Then the diagonalization of \(\Sigma\) is
\[
\Sigma = TVDV'T' + \tau^2 I.
\]
The eigenvectors of the largest eigenvalues of \(\Sigma\) are the same as those of \(TVDV'T'\), and these are therefore given by \(U = TV\). Hence \(P = TVV'T'\).

**Lemma 3** If \(q \leq \text{rank}(G)\), as \(p \to \infty\),
\[
\mathbb{E} \left[ \|\hat{P} - P\|^2 \mid F \right] = O(p^{-1}).
\]
almost surely. Furthermore, if \(q = m\) or if the \(q\)-th and \(q + 1\)-th eigenvalues of \(G\) are distinct, the statement holds uniformly in \(n\).

This lemma is a generalization of Lemma 2 in Van Houwelingen (1984).

**Proof:** By definition \(\hat{P}\) maximizes \(\text{trace}(SP)\) among all projection matrices of rank \(q\). Call \(R = S - \Sigma\). We have
\[
\text{trace}\{\Sigma(P - \hat{P})\} = \text{trace}\{SP(P - \hat{P})\} + \text{trace}\{(I - P)\Sigma(I - P)(P - \hat{P})\} \\
\leq \|R\| \cdot \|P - \hat{P}\|,
\]
the last statement being an application of the Schwartz inequality. Let \(\lambda_q\) and \(\lambda_{q+1}\) be the \(q\)-th and \(q + 1\)-th largest eigenvalues of \(\Sigma\). Then
\[
\text{trace}\{\Sigma(P - \hat{P})\} = \text{trace}\{SP(P - \hat{P})\} + \text{trace}\{(I - P)\Sigma(I - P)(P - \hat{P})\} \\
= \text{trace}\{(I - \hat{P})P\Sigma(I - P)\hat{P}\} - \text{trace}\{(I - P)\Sigma(I - P)\hat{P}\} \\
\geq \lambda_q \text{trace}(P - P\hat{P}) - \lambda_{q+1} \text{trace}(\hat{P} - P\hat{P}) \\
= \frac{1}{2}(\lambda_q - \lambda_{q+1})\|P - \hat{P}\|^2.
\]
The final equation uses \(\text{trace}\{(P - \hat{P})^2\} = \text{trace}(P) - 2\text{trace}(P\hat{P}) + \text{trace}(\hat{P})\) and \(\text{trace}(P) = \text{trace}(\hat{P})\). Combining (6.12) and (6.13) yields
\[
\|P - \hat{P}\| \leq \frac{2}{\lambda_q - \lambda_{q+1}} \|R\|.
\]

By the randomness of \(F\) and the assumption that \(q \leq \text{rank}(G)\), the first \(q + 1\) eigenvalues of \(\Sigma\) are almost surely distinct, so \(\lambda_q - \lambda_{q+1} > 0\) and the first
statement of this lemma for fixed $n$ follows directly from Lemma 1 by squaring and taking expectations.

To prove the uniformity we again remark that $n^{-1}F'F \to I$ (a.s.) as $n \to \infty$. Hence $n^{-1}(F'F)^{1/2}G(F'F)^{1/2} \to G$ and consequently $n^{-1}(\lambda_q - \lambda_{q+1})$ tends to a non-zero limit almost surely, by the assumption on the eigenvalues of $G$. Therefore the upper bound

$$\frac{2n}{\lambda_q - \lambda_{q+1}} \|n^{-1}R\|.$$ remains bounded in $n$ almost surely. □

**Lemma 4** If $q \leq \text{rank}(G)$, there is a $q \times q$ rotation matrix $W$, depending on $U$ and $\hat{U}$, such that as $p \to \infty$,

$$E[\|\hat{U} - UW\|^2 | F] = O(p^{-1})$$

almost surely. Furthermore, if $q = m$ or if the $q$-th and $q + 1$-th eigenvalue of $G$ are distinct, the statement holds uniformly in $n$.

**Proof:** First remark that $\hat{P}P$ almost surely has rank $q$, so that $U'\hat{U}$ also has rank $q$ and is invertible. Choose $W = U'\hat{U}((U'UU')^{1/2})^{-1}$, which is a $q \times q$ orthogonal matrix. Then

$$\|\hat{U} - UW\|^2 = \text{trace}(U'UU') - 2\text{trace}(UW\hat{U}') + \text{trace}(UU')
= \text{trace}(P) - 2\text{trace}\{(UU')^{1/2}\} + \text{trace}(\hat{P}).$$

The eigenvalues of $(U'UU')^{1/2}$ are the singular values of the matrix $P\hat{P}$ and therefore the square roots of the eigenvalues of the matrix $P\hat{P}P$. The eigenvalues of the latter matrix are between zero and one, so the square roots of these eigenvalues are larger than the eigenvalues themselves. Hence

$$\text{trace}\{(UU')^{1/2}\} \geq \text{trace}(U'UU') = \text{trace}(P\hat{P}),$$
so that

$$\|\hat{U} - UW\|^2 \leq \text{trace}(P) - 2\text{trace}(P\hat{P}) + \text{trace}(\hat{P})
= \|P - \hat{P}\|^2.$$ The statements of the lemma now follow immediately from their counterparts in Lemma 3. □
Proof of Theorem 3: First we apply Lemma 4 both to the matrix \( \tilde{U} \). There is an \( m \times m \) rotation matrix \( \tilde{W} \) such that \( \mathbb{E}[\| \tilde{U} - U_m \tilde{W} \|^2 | F] = O(p^{-1}) \), where \( U_m \) is (a rotation of) the \( n \times m \) semi-orthogonal matrix of the eigenvectors of the \( m \) largest eigenvalues of \( \tilde{\Sigma} = F\tilde{G}F' + \tau^2 I \). As \( \tilde{G} \) is full rank, we can take \( U_m = F (F'F)^{-1/2} \).

Next we apply Lemma 4 to the matrix \( \hat{U} \). There is a \( q \times q \) matrix \( W \) such that \( \mathbb{E}[\| \hat{U} - U_q W \|^2 | F] = O(p^{-1}) \), where \( U_q \) is (a rotation of) the \( n \times q \) semi-orthogonal matrix of the eigenvalues of the \( q \) largest eigenvalues of \( \hat{\Sigma} = F\tilde{G}F' + \tau^2 I \). As the matrix of the eigenvectors of the \( m \) largest eigenvalues is a rotation of \( U_m \), the matrix of the eigenvectors of the \( q \) largest eigenvalues is a \( U_m V_0 \) for some \( m \times q \) semi-orthogonal matrix \( V_0 \).

Remark that \( \tilde{\Lambda} = \tilde{U}'\tilde{S}\tilde{U} < \infty \) as \( p \to \infty \), so it is easily checked that \( \hat{Y} = I + O_p(p^{-1}) \) uniformly in \( n \).

Define \( V = \tilde{W}V_0 W \). Then

\[
\frac{1}{p} \mathbb{E}[\| \hat{A} - V'\hat{A} \|^2 | F] = \frac{1}{np} \mathbb{E}[\| (\hat{U}' - V'\hat{Y}^{1/2}\hat{U}')X \|^2 | F] 
\leq \frac{1}{np} \mathbb{E}[\| X \|^2 | F] \cdot \mathbb{E}[\| \hat{U}' - V'\hat{Y}^{1/2}\hat{U}' \|^2 | F]
\]

by the Schwartz inequality. The first factor on the right hand side of the inequality converges to a finite limit as \( p \to \infty \), a.s. uniformly in \( n \). The second factor can be bounded further by

\[
\mathbb{E}[\| \hat{U}' - V'\hat{Y}^{1/2}\hat{U}' \|^2 | F] \leq \mathbb{E}[\| \hat{U}' - WV_0U_m' \|^2 | F] + \mathbb{E}[\| V'\hat{Y}^{1/2}\hat{U}' - WV_0U_m' \|^2 | F]
\]

Both terms are \( O(p^{-1}) \) (a.s.) uniformly in \( n \) by Lemma 4, which proves the statement about \( \hat{A} \) and \( \hat{A} \). The proof of the statement about \( \hat{\beta} \) and \( \hat{\beta} \) is completely analogous.

\[\square\]

**Theorem 4**

We first formulate and prove a lemma on \( \hat{\tau}^2 \)

**Lemma 5** If \( r > m \), as \( p \to \infty \),

\[
\mathbb{E}[(\hat{\tau}^2 - \tau^2)^2 | F] = O(p^{-1})
\]

almost surely and uniformly in \( n \).
Proof: We have
\[
\hat{\tau}^2 - \tau^2 = \frac{1}{n-r} \text{trace}(S\hat{Q}) - \frac{1}{n-r} \text{trace}(\tau^2 \hat{Q}) = \frac{1}{n-r} \text{trace}\{\hat{Q}(S - \Sigma)\} + \frac{1}{n-r} \text{trace}\{\hat{Q}(\Sigma - \tau^2 I)\}.
\]

Construct \(\hat{P}\) and \(P\) from \(S\) and \(\Sigma\) for \(q = \text{rank}(G)\). Then \(P(\Sigma - \tau^2 I) = \Sigma - \tau^2 I\) and \(\text{trace}(\hat{Q}T) \leq \text{trace}\{(I - \hat{P})T\}\) for any positive semi-definite \(T\). We have
\[
|\hat{\tau}^2 - \tau^2| \leq \frac{1}{n-r} |\text{trace}(S - \Sigma)| + \frac{1}{n-r} \text{trace}\{(I - \hat{P})P(\Sigma - \tau^2 I)\} \leq \frac{1}{n-r} \|S - \Sigma\| + \frac{1}{n-r} \|\Sigma - \tau^2 I\|.
\]

The result follows when we remark that \(\|P - \hat{P}\|^2 = \frac{1}{2} \|P - \hat{P}\|^2\) and apply Lemmas 1 and 3. 

Proof of Theorem 4: Write \(x_{\text{new}} = Af_{\text{new}} + e_{\text{new}}\) and \(X = FA + E\). We have
\[
\hat{\gamma}'x_{\text{new}} = y'\hat{U}(\hat{U}'X\Theta X'\hat{U} + p\hat{\tau}^2 I)^{-1}\hat{U}'X\Theta Af_{\text{new}} + \hat{\gamma}e_{\text{new}}
\]
\[
= \frac{1}{p} y'\hat{P}(S - \tau^2 I)^{-1}\hat{P}FA\Theta Af_{\text{new}} + \frac{1}{p} y'\hat{P}(S - \tau^2 I)^{-1}\hat{P}E\Theta Af_{\text{new}} + \frac{1}{p} y'\hat{P}(S - \tau^2 I)^{-1}\hat{P}X\Theta e_{\text{new}},
\]
which is a sum of three complicated-looking terms, which we shorthand \(t_1\) (6.14), \(t_2\) (6.15) and \(t_3\) (6.16) in the order they appear above. We study the behaviour of the three terms when \(p \to \infty\). The calculations are tedious but straightforward. They mainly involve repeated application of Lemmas 1 and 3.

We have
\[
\mathbb{E}[(t_1 - \tilde{y}_{\text{new}})^2 | F] = O(p^{-1})
\]
\[
\mathbb{E}[t_2^2 | F] = O(p^{-1})
\]
\[
\mathbb{E}[t_3^2 | F] = O(p^{-1}),
\]
all uniformly in \(n\) under the conditions given. This proves the theorem. 
