# INVERSE PREDICTION

Philip J. Brown
University of Kent at Canterbury, UK

Inverse prediction concerns predicting one set of measurements from another set when the direction of causation and possibly error operates in the opposite direction. An example is that of calibration in which it is desired to use a cheap and quick but error prone measurement ($Y$) to predict the true amount of a constituent, ($x$), which in itself can be measured accurately in laboratory conditions but at much greater effort or cost. After taking data on samples using both measurements (the training or learning data) it is desired to use the cheap measurement in future.

The paradigm extends to multivariate response $\boldsymbol{Y}$ and composition vector $\mathbf{x}$. Relationships may be linear or non-linear. Applications range from simple instrument calibration to the use of spectroscopic (multivariate) responses for prediction of composition, to image reconstruction from noisy data. An example that will serve to focus ideas is given by Karlsson *et al* (1995) on the determination of nitrate in municipal waste water by UV spectroscopy, and discussed in the review of multivariate calibration by Sundberg (1999).

If parametric statistical analysis is employed then there are well developed estimation procedures, be they classical sampling, likelihood based, or the increasingly influential Bayesian approach. In this article we will review these. We will also comment on the use of the prediction of $x$ directly and the symmetric approach to $x, y$ embodied in latent variable methods.

Whatever statistical approach is adopted there is a need to keep the particular features of the application clearly in focus. Some preprocessing of the data may be necessary to reduce unwanted confounding. It may be that special features need to be modelled. Graphical diagnostics are invaluable in detecting important departures from assumptions. For an earlier discussion of some of these issues see (Martens and Næs, 1989).

## 1 Simple linear calibration

Suppose we have a training sample of $n$ observations $(x_i, y_i), i = 1, \dots, n$. In *controlled* calibration the data would have arisen from a designed set of $x-$values or standards.

$$Y_i = \alpha + \beta x_i + \epsilon_i \tag{1}$$

Here the $\epsilon_i$ are independent and identically normally distributed with zero mean and variance $\sigma^2$. In *random* calibration the set of $x-$ values will be a sample of those generally encountered. Hybrid designs might result from sub-sampling purposely such a random selection, perhaps to increase the variance by thinning out central values so as to increase the accuracy of estimation in the conditional distribution of $Y$ given $x$. Some more extreme $x$ values might even be added to the sample to extend the range of validity.

The statistician's standard approach to controlled calibration has a long pedigree stemming from Eisenhart (1939), and is to regress $Y$ on $x$ and then invert the estimated relationship to predict $x$ in future. If $x$ is random or the $x-$values can be safely assumed to be typical of those likely to arise in future, then to predict $x$ it is sensible to regress $X$, random, on $y$, treated as conditionally fixed. Under an extension of the Gauss- Markov theorem for quadratic prediction loss (Brown, 1993, section 3.3), this is the best linear unbiased predictor. It loses this optimality when there is one future $x$ to predict but $m$ replicates in $Y$ at this future value, as is explored below.

Let us complete the model for the training data with a similar one for the prediction data in the linear model for controlled $x$. Imagine a single unknown $x$ to predict. We denote it by $\xi$ to emphasize that it is unknown and needs to be predicted. At this value of $\xi$ we have observed $m$ replicates of $Y$. To further emphasize that these $Y$ values are for prediction and may just be different from the training data $Y$ carefully obtained under laboratory conditions, we use $Z$ in place of $Y$. The prediction model is then

$$Z_j = \alpha + \beta\xi + \epsilon_j,\ j = 1, \ldots, m. \tag{2}$$

Here it is generally assumed that the variance of the normally distributed error $\epsilon_j$ is $\sigma^2$, identically distributed to errors in (1). Departures from these assumptions would include the possibility of a larger error variance than $\sigma^2$, and correlated replicates. We have already indicated that such departures are worth checking. The classical estimator from the regression of $Y$ or $Z$ on $x$ or $\xi$ is

$$\hat{\xi}^C = (\bar{Z} - \hat{\alpha})/\hat{\beta} \tag{3}$$

where $\bar{Z} = \sum_1^m Z_j/m$ and least squares estimates $\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x},\ \hat{\beta} = s_{xy}/s_{xx}$, with $s_{xy} = \sum_1^n (x_i - \bar{x})(y_i - \bar{y}),\ s_{xx} = \sum_1^n (x_i - \bar{x})^2$. It terms of departures from training means eqn (3) becomes $(\hat{\xi}^C - \bar{x}) = (\bar{Z} - \bar{y})/\hat{\beta}$. The inverse predictor for $m = 1$ from the regression of $X$ on $y$ is simply

$$\hat{\xi}^I = \hat{\alpha}^* + \hat{\beta}^* Z \tag{4}$$

or in terms of departures from training means, $(\hat{\xi}^I - \bar{x}) = \hat{\beta}^*(Z - \bar{y})$ with least squares estimates $\hat{\alpha}^* = \bar{x} - \hat{\beta}^*\bar{y},\ \hat{\beta}^* = s_{xy}/s_{yy}$ and $s_{yy} = \sum_1^n (y_i - \bar{y})^2$.

The two regression lines are related by

$$(\hat{\xi}^I - \bar{x}) = r^2(\hat{\xi}^C - \bar{x}) \tag{5}$$

where $r^2 = \frac{s_{xy}^2}{s_{xx}s_{yy}}$ the squared correlation. Since $r^2 \leq 1$, the inverse estimator is closer to the mean of $x$ than the classical estimator. With a perfect correlation of unity the estimators will coincide, otherwise the difference between the two depends on the strength of the relationship through $r^2$.

The shrinkage of the inverse estimator towards the mean relative to the classical estimator is reflected in the assumption of a common normal distribution for the $x'$s and $\xi$ in alternative derivations of the inverse estimator. All approaches to inference which properly account for the uncertainty use the two sources of uncertainty, in the training data through model (1) with both $x$ and $y$ data and through the future observation(s) with just $Z$ data in model (2). The Bayes approach to this amalgamation stems from Hoadley (1970) and was extended by Brown (1982) to the multivariate general linear model. Broadly it uses the training data to derive the posterior distribution of the regression parameters and then integrates these out from the prediction model for $Z$ to form a predictive density of the only remaining unknown, $\xi$, which combines with the prior for $\xi$ to form the posterior. When there are prediction replicates these will add information (residual sum of squares $s'^2$) about $\sigma^2$, provided this can be safely assumed to be common to both prediction and training. In this case the typical form of posterior is

$$\pi(\xi|\boldsymbol{Z}, \boldsymbol{Y}, \boldsymbol{X}\ ) \propto f(\bar{Z}|\boldsymbol{Y}, \boldsymbol{X}, s'^2, \xi)\pi(\xi|\boldsymbol{X}\ ),$$

where for detailed assumptions see Brown (1993), chapter 5.

Here however we will tie together approaches to univariate, multivariate and nonlinear models through the profile likelihood of $\xi$. This uses maximisation of likelihood rather than the integration of the Bayes approach. It is formed by maximising over parameters $\alpha, \beta, \sigma^2$ the likelihood formed by combining the $n$ observations $(x_i, Y_i), i = 1, \ldots, n$ and $m$ 'observations' $(\xi, Z_j)$, $j = 1, \ldots, m$ for fixed $\xi$. The profile or relative likelihood (Kalbfleisch and Sprott, 1970) gives maximum likelihood estimate of $\xi$ if it is further maximized over $\xi$, and the profile likelihood is often standardized to having an overall maximum of unity at the MLE. The profile likelihood itself gives the relative likelihood of different values of $\xi$. By taking all values of $\xi$ such that this profile likelihood is greater than some constant we generate likelihood ratio confidence intervals. For the controlled calibration situation the profile likelihood is proportional to

$$\underset{\alpha, \beta, \sigma^2}{max}\ [\prod_1^n f(y_i \mid x_i, \alpha, \beta, \sigma^2) \prod_1^m f(z_j \mid \xi, \alpha, \beta, \sigma^2)].$$

The explicit form is given later for the more general multivariate case in formula (10). This profile likelihood as a function of $\xi$ has a maximum at $\hat{\xi}^C$ given by (3), verifying its maximum likelihood credentials. However if one assumes some of the $x-$values and $\xi$ are random normal, which one is free to do without affecting the distribution of $Y \mid x$, where all have the same unknown mean and variance, then a vital link between the unknown $\xi$ and the $x-$sample is established. The profile likelihood augmented by the normal $x_i, \xi$ data now shifts its peak towards the mean of the random $\xi$. In the case where all the $x-$values *and* $\xi$ form a $N(\mu, \tau^2)$ random sample, then the profile likelihood is the controlled case relative likelihood multiplied by

$$\underset{\mu, \tau^2}{max}\ [\prod_1^{n+1} f(x_i \mid \mu, \tau^2)]$$

where $x_{n+1} = \xi$. This extra term is like the kernel of a Student distribution, $c_{1,n}^2 = [1 + 1/n + (\xi - \bar{x})^2/s_{xx}]^{-\frac{(n+1)}{2}}$ where $\bar{x} = \sum_1^n x_i/n$. Without prediction replicates, $m = 1$, the combined term is maximized by $\hat{\xi}^I$ given by (4) (Brown, 1993, Section 2.5). When there are prediction replicates the maximum of the profile likelihood, let us call it $\hat{\xi}^{I^*}$, is not equal to $\hat{\xi}^I$ extended by $\bar{Z}$ replacing $Z$. The predictor $\hat{\alpha}^* + \hat{\beta}^* \bar{Z}$ is biased towards $\bar{x}$, as noted for example by Srivastava (1995) in simulations comparing classical and inverse predictors in multi-univariate linear calibration. A series expansion of the log-profile likelihood as a quadratic in $\xi$ gives a maximising value and approximate predictor $(\hat{\xi}^{I^*} - \bar{x}) = k\hat{\beta}^*(\bar{Z} - \bar{y})$ with expansion factor $k$ given as

$$k = \frac{m\hat{\beta}^2 + m[s^2/s_{xx}]}{m\hat{\beta}^2 + [\frac{s_+^2}{(n+m)}\frac{(n+1)}{s_{xx}}]}, \tag{6}$$

with $s^2$, $s_+^2$ residual sums of squares from the training data, and training data augmented with mean corrected sums of squares from the replicates, respectively. We reiterate that it is important here to check for departures from assumptions especially differences between error variances in training and prediction, and correlated replicates.

The factor $k$ given by (6) is unity with no prediction replicates and is monotone increasing in $m$, with a limiting value as $m \to \infty$ of $1 + \frac{s^2/s_{xx}}{\hat{\beta}^2} = 1/r^2$, when as a consequence of (5) the estimator reverts to $\hat{\xi}^C$. Similar relationships between the estimators may be explored by adopting a Bayes approach, see Brown (1982) for the multivariate general linear model case. In the published discussion of this paper (i) T. Fearn emphasizes that the design for calibration will typically choose or sub-select $x$'s that span the range of likely future values of $\xi$, (ii) J. Copas emphasises the essential inferential need for a link between $\xi$ and the training $x$-sample which underpins inverse estimation.

Likelihood ratio confidence intervals may be readily specified. For controlled calibration the profile likelihood is a monotone function of the pivotal statistic

$$[\bar{Z} - \bar{y} - \hat{\beta}(\xi - \bar{x})]/[\hat{\sigma}(1/m + 1/n + \xi^2/s_{xx})^{1/2}] \tag{7}$$

with $(n + m - 3)\hat{\sigma}^2 = s_+^2$, and has a Student t-distribution on $(n + m - 3)$ degrees of freedom. Notice here that uncertainty in $\xi$ persists even if $n \to \infty$ when the regression parameters become effectively known: the $\hat{\sigma}/\sqrt{m}$ term in the denominator of (7) remains and only reduces to zero as $m \to \infty$.

In the inverse estimation context, a corresponding Student sampling or Bayes predictive distribution from the regression of $X$ on $y$ is readily available. It forms the basis for confidence intervals when there are no prediction replicates, $(m = 1)$, and may be extended through the profile likelihood (or Bayes posterior distribution) for prediction with replicates, $m > 1$.

# 2 Multivariate Calibration

Simple single variable versus single variable calibration as featured above provides insights into the essential differences between classical and inverse calibration. It is however the case that most modern instruments are multivariate in $\boldsymbol{Y}$ and often multiple in $\boldsymbol{x}$. So called second-order or hyphenated instruments go one step further and deal with matrix $\boldsymbol{Y}$ for each $\boldsymbol{x}$, (Bro, 1996; Smilde, 1997; Linder and Sundberg, 1998).

In the multivariate general linear model, model (1) with prediction model (2) may be extended to

$$\boldsymbol{Y} = \boldsymbol{1}\boldsymbol{\alpha}' + \boldsymbol{\mathcal{X}}\boldsymbol{B} + \boldsymbol{E}, \tag{8}$$

$$\boldsymbol{Z} = \boldsymbol{1}\boldsymbol{\alpha}' + \boldsymbol{1}\boldsymbol{\xi}'\boldsymbol{B} + \boldsymbol{E}^*, \tag{9}$$

where $\boldsymbol{Y}(n \times q)$, $\boldsymbol{E}\ (n \times q)$, $\boldsymbol{Z}(m \times q)$, $\boldsymbol{E}^*(m \times q)$ are random matrices, $\boldsymbol{\mathcal{X}}(n \times p)$ is a matrix of fixed constants as are the vectors of units, $\boldsymbol{1}$,which are respectively $n \times 1$ and $m \times 1$; $\boldsymbol{\xi}$is a $p \times 1$ vector of $p$ unknowns. Rows of the errors $\boldsymbol{E}$ and $\boldsymbol{E}*$ are independent multivariate normal with mean the zero vector and covariance matrix the the $q \times q$ unknown matrix $\boldsymbol{\Gamma}$. This model was developed in (Brown, 1982). In the context of the municipal waste water example we have $n = 125$, observations, $q = 316$ UV-absorbances, and there is just one component (nitrate) , $p = 1$, of waste water of interest.

The profile likelihood for $\boldsymbol{\xi}$ from (8),(9) is given as

$$[c^2_{m,n}(\boldsymbol{\xi})/\{c^2_{m,n}(\boldsymbol{\xi}) + [\bar{\boldsymbol{Z}} - \bar{\boldsymbol{y}} - \hat{\boldsymbol{B}}'(\boldsymbol{\xi} - \bar{\boldsymbol{x}})]'\boldsymbol{S}_+^{-1}[\bar{\boldsymbol{Z}} - \bar{\boldsymbol{y}} - \hat{\boldsymbol{B}}'(\boldsymbol{\xi} - \bar{\boldsymbol{x}})\}]\}]^{\frac{n+m}{2}}. \tag{10}$$

Here $\boldsymbol{S}_+$ is the $q \times q$ residual sum of products matrix augmented by $m-1$ replicate information in (9), and

$$c^2_{m,n}(\boldsymbol{\xi}\ ) = [1/m + 1/n + (\boldsymbol{\xi} - \bar{\boldsymbol{x}})'\boldsymbol{S}_{\boldsymbol{XX}}^{-1}(\boldsymbol{\xi} - \bar{\boldsymbol{x}})] \tag{11}$$

with $\boldsymbol{S}_{\boldsymbol{XX}}$ the $p \times p$ sum of mean corrected products matrix. The simple estimator analogous to (3) is the generalized least squares estimator

$$\hat{\boldsymbol{\xi}}^{\boldsymbol{GLS}} - \bar{\boldsymbol{x}} = (\hat{\boldsymbol{B}}\boldsymbol{S}_+^{-1}\hat{\boldsymbol{B}}')\hat{\boldsymbol{B}}\boldsymbol{S}_+^{-1}(\bar{\boldsymbol{Z}} - \bar{\boldsymbol{y}})$$

This does not maximize (10) unless $p = q$, (Brown and Sundberg, 1987). When there are relatively few observations, $\boldsymbol{S}_+$ is not of full rank and generalized inverse versions are possible if $n \leq p + q$, but are somewhat unstable, see Sundberg (1996) who compares the precision of a range of estimators and discusses putting reduced parametric factor structure into $\boldsymbol{\Gamma}$ or Bayesian forms of regularization or shrinkage.
(See also the EofE entry on SHRINKAGE REGRESSION.)
Another strategy revolves around response selection, see for example Spiegelman *et al* (1996). Variable selection is usually easiest to implement in the regression of $\boldsymbol{X}$ on $\boldsymbol{Y}$; see (Brown *et al.*, 1998) for fully multivariate Bayesian stochastic search variable selection employing Markov chain Monte Carlo.

For a review of confidence regions in controlled calibration see Oman (1999). As with the univariate case, uncertainty in $\boldsymbol{\xi}$ can only go to zero as both training sample size $n$, and prediction replicates $m$ both become large. When as is usual the number of responses $q$ is much greater than the number of components, $p$, of $\boldsymbol{\xi}$ to be predicted then conflict can arise whereby some components of the future $\boldsymbol{Y}$ suggest a quite different value of $\boldsymbol{\xi}$ to other components: there is no sufficiency reduction to iron this out due to the curved nature of the exponential family, and likelihood based confidence regions will reflect this conflict by becoming flatter.

Prediction from the regression of $\boldsymbol{X}$ on $\boldsymbol{Y}$ may be formulated in a similar way to the univariate profile likelihood approach. The profile likelihood formed by augmenting models (8),(9) with the assumption of some random $p-$vectors $\boldsymbol{X_i}$, taken to the limit of all $n$ random, modifies (10) by a factor of $c^2_{1,n}\ (\boldsymbol{\xi})^{-\frac{n+1}{2}}$, a special case of (11), and reverts to the predictive likelihood from the regression of $\boldsymbol{X}$ on $\boldsymbol{Y}$. Confidence procedures are just obtained from multivariate predictive Student distributions.

The regressions of $\boldsymbol{X}$ on $\boldsymbol{Y}$ and $\boldsymbol{Y}$ on $\boldsymbol{X}$ lead to $\boldsymbol{\xi}$ predictors that after a suitable transformation are simply proportional with the squared canonical correlations as proportionality factors, (Brown, 1982), extending the simple univariate relationship (5). Chemometrician's have often preferred to formulate models symmetrically in $\boldsymbol{X}$ and $\boldsymbol{Y}$, and have eschewed statistician's focus on random and controlled calibration in favour of algorithmic methods such as partial least squares, see particularly the discussion of Sundberg's paper (1999) by H. Martens and and S. Wold whose arguments are persuasive but also see Sundberg's response. For a general latent variable approach see (Burnham *et al.*, 1999).

# 3 Non-linear calibration

New features arise if the relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$ is non-linear even in the univariate case. Let us focus on controlled calibration. In the case of multivariate $\boldsymbol{Y}$ it may be possible to select a subset of $\boldsymbol{Y}$ variables for which the relationship is linear without discarding too much information (Brown, 1993, chapter 7). If we do not have a surplus of variables and want to retain all, then two situations may be distinguished.

(i) The regression of $\boldsymbol{Y}$ on $\boldsymbol{x}$ is linear in the parameters but non-linear in $\boldsymbol{x}$.

(ii) The regression of $\boldsymbol{Y}$ on $\boldsymbol{x}$ is non-linear in the parameters as well as $\boldsymbol{x}$.

In the first case although the linear model applies to the calibrating data, prediction is still non-linear. An example is polynomial regression. In the second case standard inferential methods still apply (likelihood and Bayes) but linearity in the model parameters cannot be exploited. For case (i), the profile likelihood is still given by (10), (11) if we replace $\boldsymbol{\xi}$ by $h(\boldsymbol{\xi})$ a $p \times 1$ vector of $p^* < p$ components in $\boldsymbol{\xi}$. Lundberg and de Maré (1980) develop the univariate polynomial $x$ case and note that if the mean is monotone in $x$ over the range of $x$ and possible $\xi$, then a single unique consistent estimate $\hat{\xi}$ exists, which is indeed the MLE. Their asymptotic confidence region is just obtained by linearising the mean function about this estimate. In general provided the $\boldsymbol{x}-$relationship is not too curved, linearization about the maximum likelihood value of $\boldsymbol{\xi}$ can provide straightforward confidence regions, see also (Oman, 1999).

Clarke (1992) provides a profile likelihood solution for case (ii) with a single unknown component of $x$, and mean functions monotone in $x$. As with the earlier discussion this just involves writing down the likelihood function from both training and prediction for given $\xi$ and maximising over the other unknown model parameters. Simple confidence procedures are available when approximating by linearization of the mean function both in $\xi$ and the model parameters about their maximum likelihood estimates. There remains the possibility of conflict between the estimates of $\xi$ on different response axes, (Sundberg, 1999). The example given of rhinoceros data with 2 responses (anterior and posterior horn lengths) and a single $x$ (age) could benefit from injecting the exchangeability assumption of $\xi$ and the $x$-sample. It is analysed by Bayesian simulation techniques by (du Plessis and van der Merve, 1996).

# References


Bro, R. (1996). Multiway calibration. Multilinear PLS. *Journal of Chemometrics*, **10,** 47–61.

Brown, P. J. (1982). Multivariate calibration (with discussion). *Journal of the Royal Statistical Society,* B, **44,** 287–321.

Brown, P. J. (1993). *Measurement, Regression,and Calibration.* Oxford: Clarendon Press.

Brown, P. J. and Sundberg, R. (1987). Confidence and conflict in multivariate calibration. *Journal of the Royal Statistical Society,* B, **49,** 46–57.

Brown, P. J., Vannucci, M. and Fearn, T. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series* B, **60,** 627–641.

Burnham, A. J., MacGregor, J. F. and Viveros, R. (1999). A statistical framework for multivariate latent variable methods based on maximum likelihood. *Journal of Chemometrics*, **13,** 49–65.

Clarke, G. P. Y. (1992). Inverse estimates from a multiresponse model. *Biometrics*, **48,** 1081–1094.

du Plessis, J. L. and van der Merve, A. J. (1996). Bayesian calibration in the estimation of the age of rhinoceros. *Annals of the Institute of Statistical Mathematics*, **48,** 17–28.

Eisenhart, C. (1939). The interpretation of certain regression methods and their use in biological and industrial research. *Annals of Mathematical Statistics*, **10,** 162–186.

Hoadley, B. (1970). A Bayesian look at inverse regression. *Journal of the American Statistical Association*, **65,** 356–369.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large numbers of parameters. *Journal of the Royal Statistical Society,* B, **32,** 175–208.

Karlsson, M., Karlberg, B. and Olsson, R. J. O. (1995). Determination of nitrate in municipal waste water by UV spectroscopy. *Analytica Chemica Acta*, **312,** 107–113.

Linder, M. and Sundberg, R. (1998). Second order calibration: bilinear least squares regression and a simple alternative. *Chemometrics and Intelligent Laboratory Systems*, **42,** 159–178.

Lundberg, E. and de Maré, J. (1980). Interval estimates in the spectroscopic calibration problem. *Scandinavian Journal of Statistics*, **7,** 40–42.

Martens, H. and Næs, T. (1989). *Multivariate calibration.* Wiley, Chichester.

Oman, S. D. (1999). Multivariate calibration. In *Multivariate analysis, design of experiments, and survey sampling* (ed. S. Ghosh), pp. 265–299. Marcel Dekker: New York, Basel.

Smilde, A. K. (1997). Comments on multilinear PLS. *Journal of Chemometrics*, **11,** 367–377.

Spiegelman, C., Wang, S. and Denham, M. (1996). Asymptotic minimax calibration estimates. *Chemometrics and Intelligent Laboratory Systems*, **32,** 257–263.

Srivastava, M. S. (1995). Comparison of the inverse and classical estimators in multi-univariate linear calibration. *Communications in Statistics*, **24,** 2753–2767.

Sundberg, R. (1996). The precision of the estimated generalized least squares estimator in multivariate calibration. *Scandinavian Journal of Statistics*, **23,** 257–274.

Sundberg, R. (1999). Multivariate calibration– direct and indirect regression methodology. *Scandinavian Journal of Statistics*, **26,** 161–207.