Linear Multiresponse Models
-Theoretical Developments and Applications in Porcine

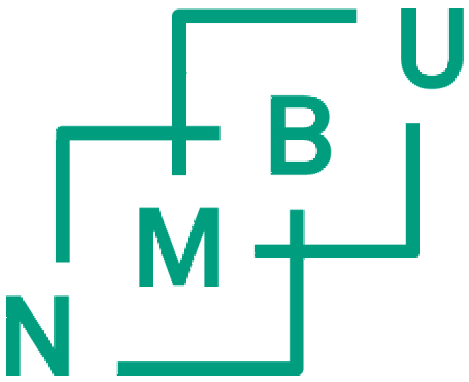
Lineære multiresponsmodeller
-Teoretiske nyvinninger og praktiske anvendelser for svin

Philosophiae Doctor (PhD) Thesis

Lars Erik Gangsei

Department of Chemistry, Biotechnology and Food Science
Faculty of Environmental Science and Technology
Norwegian University of Life Sciences

Ås (2016)



Thesis number 2016:34
ISSN 1894-6402
ISBN 978-82-575-1361-0

Preface

Formalia

The research for and writing of this thesis were completed in the period 2013–2016 at the Department of Chemistry, Biotechnology and Food Science (IKBM) at the Norwegian University of Life Sciences (NMBU) under direction of main supervisor Solve Sæbø, and co–supervisors Trygve Almøy, Ole Alvseike and Jørgen Kongsro. I have been employed by Animalia, where I have been affiliated with the classification group headed by Morten Røe.

Thesis structure

The five papers that constitute the basis for the thesis can be divided into two major, related themes. Papers I–III deal with linear regression containing a bivariate response variable with missing data. Papers IV and V deal with 3D image analysis applied to computed tomography (CT) images of Norwegian pigs. The papers are reproduced in full in the last part of this thesis.

The introduction of this thesis is organized in two chapters reflecting the two major themes. The first chapter deals with the missing data problem and the "empirical Bayes machinery" which fit well with papers I–III. This chapter also contains reviews and summaries of papers I–III.

The second chapter deals with *in vivo* CT scans of pigs. It aims at explaining the basic principles of atlas segmentation, and how papers IV and V fit into these principles. In addition, summaries of paper IV and V are included, together with a closing section in which I describe how the principles of atlas segmentation might be incorporated into the "empirical Bayes machinery".

Notations

Some mathematical formulas are inevitable, including in the introduction. With a couple of exceptions, the notations are consistent throughout the thesis. Scalars are denoted by lowercase italic characters, vectors by lowercase bold italic characters and matrices by uppercase bold italic characters. The single elements of any matrix are denoted by the corresponding lowercase italic letter and a subindex, i.e. w_{ij} is the element of the ith row and jth column of $\boldsymbol{W}$. In general, Greek letters are used for parameters and Latin letters are used for random variables. The risk function, $R_{\hat{\boldsymbol{\theta}}}$, for an estimator or predictor, $\hat{\boldsymbol{\theta}}$, is the expectation of the standard quadratic loss function. I.e. the risk is given by the expected sum of squares for the difference between the estimator, and its true value, $\boldsymbol{\theta}$.

Acknowledgements

It is a somewhat puzzling, although not coincidental, situation that my main supervisor, Solve Sæbø, and I both finished Master degrees at NMBU in 1999. We studied different programs: Solve was a statistician already then; I studied biology, but we participated in a couple of courses simultaneously. Although I had a rewarding job after finishing my studies, I regretted having studied biology in the first place. This led me to first take a master's degree and then this PhD, both in applied statistics, and with Solve as my main supervisor. I can vividly imagine how challenging it must be to supervise a former fellow student, especially an unmarried, headstrong and "old" PhD-student like me. My perception of working under Solve's supervision can be summed up through the following four phrases: Mostly free to mind my own business, always free to ask, sometimes guided in the right direction and never guided in the wrong direction.

I have two sets of skilled colleagues, the Biostatistics group at NMBU and my colleagues at Animalia, both absolutely helpful. Thanks to all members and associates. In particular, I will thank my co–supervisor Trygve Almøy for his helpfulness.

I must also mention Ole Alvseike and Morten Røe at Animalia. If Ole got a dollar for every new idea he gets, he would soon be a billionaire. On the other hand, if he got a dollar for every *good* idea that occurs to him, he would only be a millionaire. Morten likes figures. I have never seen someone so passionate about his job as Morten. Luckily, his job is figures. It is never boring, and always inspiring, to be around Ole and Morten.

My last supervisor has been Jørgen Kongsro from Topigs Norsvin. For the last two years, we have worked extensively together on the pig atlas. The help I have received from Jørgen has been excellent. I believe, and sincerely hope that our cooperation has resulted in some benefits for Jørgen and for Topigs Norsvin too.

All the prior knowledge I had of image analysis when we started working with the pig atlas was the result of a three–month study residency I had in 2013 at the Image Analysis Group at the Technical University of Denmark. I want to thank them for their hospitality and for the highly skilled teaching they provide.

Last, but not least, thanks go to my family: my parents, brothers and sister. There is a saying in Norwegian: "å være født med en sølvskje i munnen" (to be born with a silver spoon in one's mouth). As the oldest son of a nurse and a lumberjack I believe that my upbringing properly reflects this saying in a positive interpretation.

Ås, 26 February 2016

Lars Erik Gangsei

Summary

The main topic of this PhD–thesis is how to minimize the prediction error for multi–response linear regression models. Two different applications are analysed, (*i*) bivariate response with missing data and (*ii*) image analysis from computed tomography (CT). Both applications were initialized by practical problems in porcine.

We evaluated a James–Stein estimator, a biased estimator, applied to the model with missing data. As the first step, we assumed that some model parameters were known in order to reach analytic results. By assuming that the predictor variables were drawn independently from a multivariate normal distribution, we found analytical expressions for expected squared prediction error, over all training– and test–sets, as presented in paper I.

In paper II, we analyzed the same model under a Bayesian regime, omitting the assumption about known model parameters. We found a conjugate prior distribution for the unknown parameters, and showed how a particular parameterization provides an estimator with properties similar to the James–Stein estimator. Prediction of new responses might be conducted by a totally data–driven methodology, "the empirical Bayes machinery". By simulations, we showed that the prediction precision of the empirical Bayes estimator mostly was better, often with substantial margin, compared to natural competitors. In cases where competitors predicted more precisely, the margins tended to be small. In paper III, we showed that the model works for real data, i.e. for prediction of lean meat percentage in pig carcasses.

The pig breeding company Topigs Norsvin CT scans approximately 3 500 nucleus boars annually in Norway. CT provides very valuable knowledge for the breeding process, but the value would increase considerably if an effective method existed to identify various organs, commercial cuts etc. Atlas segmentation is one such method.

The atlas might be viewed as an average pig. The underlying idea is that one can transform ("squeeze") the individual pigs into the atlas–space, where organs, commercial cuts etc. are predefined. Thus, these features might be identified at the individual scale. The transformation is a multivariate linear prediction, where the coordinates in atlas–space are predicted by using basic functions of the coordinates in the individual–spaces as predictors.

For construction of the atlas, and fitting the transformations, we based our estimates on corresponding landmarks. Skeleton structure and the surface (skin) were used for identifying the landmarks. Paper IV describes solely an algorithm for automatic segmentation and identification of the major bones in the skeleton. This algorithm is essential for paper V where we describe how the atlas is constructed and used for segmentation. The algorithm is derived from basic, and well–known, image analysis techniques.

Sammendrag (*Norwegian summary*)

Hovedtemaet i denne PhD–avhandlingen er metodikk for å redusere prediksjonsfeil i lineære regresjonsmodeller med flere responsvariabler. To ulike bruksområder, (*i*) bivariat respons med manglende data og (*ii*) 3D bildeanalyse av data fra computertomografi (CT), blir behandlet. Begge har utganspunkt i praktiske problemstillinger fra svineproduksjon.

Vi analyserte en forventningsskjev estimator, kjent som James–Stein estimatoren, i problemet med manglende data. I paper I baserte vi analysene på en antagelse om at flere modellparametere var kjente størrelser. Ved å anta at forklaringsvariablene blir trukket uavhengig av hverandre fra en multivariat normalfordeling, fant vi også analytiske uttrykk for forventet kvadrert prediksjonsfeil ved bruk av ulike estimatorer.

I paper II analyserte vi modellen under et Bayesiansk regime, uten å gjøre forutsetninger om kjente modellparametere. Vi fant en konjungerende prior fordeling (conjugate prior) for de ukjente parameterene, og viste hvordan en spesiell parametrisering av denne gir en estimator med egenskaper svært like James–Stein estimatoren. Estimatoren kan beregnes ved en fullstendig data drevet metode (empirisk Bayes). Ved simuleringer viste vi at empirisk Bayes estimatoren oftest predikerer bedre, gjennomgående med betydelig margin, sammenlignet med naturlige konkurrenter. I tilfeller hvor konkurrenter er bedre er marginene små. I paper III viser vi at modellen fungerer på reelle data, prediksjon av kjøttprosent i griseslakt.

Hvert år CT scanner selskapet Topigs Norsvin ca. 3 500 norske hanngriser som er aktuelle å benytte i avlen. CT gir svært verdifull kunnskap for avlsarbeidet, men verdien vil øke betydelig dersom man finner en effektiv metode for identifisere ulike organer, stykningsdeler etc. i CT bildene. Atlas segmentering er en slik metode. Atlaset kan ses på som en gjennomsnittsgris. Den underliggende ideen er at man kan transformere ("skvise") de enkelte individene inn i atlaset (atlas–formen), hvor organer, stykningsdeler etc. er definert på forhånd. Dermed identifiserer man de ulike organene/ stykningsdelene i individet. Transformasjonen er en multivariat linær prediksjon, hvor de predikerte verdiene består av koordinater i atlas–rommet og forklaringsvariablene er basert på basis funksjoner av koordinatene i individ–rommene.

For å konstruere atlaset, og transformasjonene, baserte vi oss på landmerker (landmarks). Dette er punkter med kjente koordinater både i atlas–rommet og i individ–rommene. Skjelettstrukturen, samt overflaten (skinnet), ble benyttet for å identifisere disse punktene. Paper IV beskriver utelukkende en algoritme for automatisk segmentering og identifisering av de største knoklene i skjelettet. Denne algoritmen er viktig for paper V hvor vi beskriver hvordan atlaset konstrueres og benyttes til segmentering. Hele paper V er basert på standard bildeanalyse teknikker.

Contents

List of papers

(I) Gangsei, L. E., Almøy, T. & Sæbø, S. (2016). Theoretical evaluation of prediction error in linear regression with a bivariate response variable containing missing data. Submitted manuscript[a] to Communications in Statistics – Theory and Methods.

(II) Gangsei, L. E., Almøy, T. & Sæbø, S. (2016). Linear regression with bivariate response variable containing missing data. An empirical Bayes strategy to increase prediction precision. Submitted manuscript[a] to Communications in Statistics – Simulation and Computation.

(III) Gangsei, L. E., Kongsro, J., Olsen, E. V., Røe, M., Alvseike, O. & Sæbø, S. (2016). Prediction precision for lean meat percentage in Norwegian pig carcasses using "Hennessy grading probe 7". Evaluation of methods emphasized at exploiting additional information from Computed Tomography. Submitted manuscript[b] to Acta Agriculturae Scandinavica, Section A - Animal Science.

(IV) Gangsei, L. E. & Kongsro, J. (2016). Automatic segmentation of Computed Tomography (CT) images of domestic pig skeleton using a 3D expansion of Dijkstra's algorithm. Computers and Electronics in Agriculture, 121:191–194.

(V) Gangsei, L. E., Kongsro, J., Olstad, K., Grindflek, E. & Sæbø, S. (2016). Building an *in vivo* anatomical atlas to close the phenomic gap in animal breeding. Submitted manuscript[c] to Scientific Reports.

a: The manuscripts were submitted to the journals on 11 January 2016. We have had no feedback from the respective journals other than confirmations that they have received the manuscripts. Some minor changes are made to the manuscript in this thesis compared to the submitted manuscripts.

b: The manuscript was submitted to the journal on 12 January 2016. On 25 February 2016 we got feedback from reviewers and were offered to resubmit with minor corrections. The manuscript in this thesis has taken feedback from reviewers into account, and will, maybe with some minor changes, be resubmitted in the near future.

c: The manuscript was submitted to the journal on 22 February 2016. On 25 February 2016 we got feedback asking us for some additional information about the supplementary material and some more documentation on official approvement of the CT–experiments. We plan to do this in the near future, however the manuscript in this thesis is identical to the manuscript submitted on 22 February.

Bivariate Response – Missing Data

1 Background

1.1 Surrogate variable

The papers I–III were initiated by the practical problem dealt with in paper III, i.e. how to best predict lean meat percentage (LMP) in Norwegian pig carcasses. The prediction equation had to be fitted by a set of data from a total of 465 pig carcasses. The measurements from the optical probe "Hennasay Grading Probe 7" (HGP7), together with registrations of sex, breed and year of slaughtering, constituted the predictor variables. The LMP–S, i.e. the response variables, were fully observed using computed tomography (CT) as dissection method (Vester-Christensen et al., 2009) (LMP–CT), whereas LMP was only observed for 86 carcasses using manual dissection (Walstra and Merkus, 1995) (LMP–MD).

LMP–CT and LMP–MD are known to have a very high correlation (Daumas and Monziols, 2011), though their scale might differ. This pattern was evident for our data as the correlation between LMP–CT and LMP–MD were 0.968, with average values at 66.6 % and 60.5% respectively.

The salient point in the problem was that the LMP–S should be predicted at a scale corresponding to manual dissection. Thus we wanted to "borrow strength" from data where only LMP–CT was registered when estimating the regression parameters associated with LMP–MD, and accordingly reduce the expected prediction error for LMP–MD.

Even though our introduction to the problem in question was a practical problem involving pigs, I assume that the general problem with missing data in a response variable is very common. The fully observed response variable, i.e. LMP–CT in our case, might be viewed as a surrogate variable, which is defined as (Upton and Cook, 2014) "*a variable that can be measured (or is easy to measure) that is used in place of one that cannot be measured (or is difficult to measure)*".

Upton and Cook (2014) distinguishes between a surrogate variable and a proxy variable defined as "*a measurable variable that is used in place of a variable that cannot be measured*". A search on scholar.google.com conducted on 18 January 2016 revealed approximately 30-thousand hits for the term "proxy variable" (inside quotation marks) versus approximately 6–thousand hits for the term "surrogate variable". Without further examination, I suspect that a lot of studies using the term "proxy variable" in reality mean "surrogate variable" under the strict definition by Upton and Cook (2014).

In a number of scientific disciplines, extensive use of proxy variables and/ or surrogate variables, is a cornerstone in a large proportion of studies. I do not claim that the method described in paper II and the theory of paper I can be directly implemented to address

all such problems, at least not problems violating the assumption about independent error terms between observations, typically time-series data. However, I do claim that paper II, and in particular paper I, might shed some light on when a surrogate (response) variable might help improve the prediction precision for the primary response variable, i.e. how to deal with missing data for one response.

1.2 Model

The natural model for analysing linear regression with a bivarite response, and which is the cornerstone of papers I–III, is:

$$\boldsymbol{y}_i^T \overset{i.i.d}{\sim} N_2\left(\boldsymbol{\beta}^T\boldsymbol{x}_i^T,\ \boldsymbol{\Sigma}\right),\ i=1,\ldots,n_1; \tag{1.1}$$

where $\boldsymbol{y}_i$ denotes the ith row of $\boldsymbol{Y}_1 = [\boldsymbol{y}_1\ \boldsymbol{y}_2]$ and $\boldsymbol{x}_i$ denotes the ith row of $\boldsymbol{X}_1$, i.e. the matrix of predictor variables. The number of rows in $\boldsymbol{Y}_1$ and $\boldsymbol{X}_1$ are n_1. The n_2 first rows of $\boldsymbol{Y}_1$ and $\boldsymbol{X}_1$ are denoted $\boldsymbol{Y}_2$ and $\boldsymbol{X}_2$ and contains the observations where the response is fully observed. The $p \times 2$ matrix $\boldsymbol{\beta} = [\boldsymbol{\beta}_1\ \boldsymbol{\beta}_2]$, denotes the regression coefficients. The 2×2 matrix $\boldsymbol{\Sigma}$ denotes the error covariance matrix, with elements σ_{ij}, $i = 1, 2$, $j = 1, 2$.

1.3 Focus on prediction precision

Multiresponse linear regression is well known in the statistical literature (Mardia et al., 1979). However, when no data is missing, the uniformly minimum–variance unbiased estimator (UMVUE) for $\boldsymbol{\beta}$ equals the matrix whose columns consist of ordinary least squares (OLS) estimators for the corresponding univariate models (Srivastava, 1965). Thus the popularity of multiresponse models, at least when the ultimate goal is prediction, seems to be somewhat restricted.

In papers I–III we evaluate the prediction precision associated with different estimators for $\boldsymbol{\beta}_2$. That is, our focus is on finding an estimator that provides better prediction precision for the response with missing data than natural competitors. We do not focus on the quality of the estimates for the variance component, $\boldsymbol{\Sigma}$, or the estimates for the regression parameters associated with the response without missing data, $\boldsymbol{\beta}_1$.

The principle of focusing on a subset of model parameters or, as in our case, the prediction precision, a function for which the influence of model parameters is restricted to the estimate for $\boldsymbol{\beta}_2$, is the fundamental motivation for the focused information criterion (FIC) (Claeskens and Hjort, 2003), and model averaging based on similar principles (Hjort and Claeskens, 2003). Even though FIC relies on a frequentistic approach, and the leitmotif

in papers I–III is evaluation of an empirical Bayes estimator, the number of focal points between the two methods are abundant.

1.4 Alternative methods

We evaluated a method based on empirical Bayes principles in papers II and III. It is worth noting that the parameters in the model defined in (1.1), including the missing data challenge, might be estimated through standard methods. Examples of such methods are the expectation maximization algorithm (A. P. Dempster, 1977), within the framework of maximum likelihood estimation (MLE), see Savage (1976) for history and basic principles. The MLE estimators for variance components are known to be biased.

A related method known as restricted maximum likelihood estimation (REML), deals with the issue of biased variance components, see Harville (1977) for the basic ideas and history. A REML based algorithm, as described in Diggle et al. (2002), for estimating the parameters in (1.1) would be fairly easy to implement for the model and problem in question.

If we assume that $\sigma_{11} = \sigma_{22}$ and $\sigma_{12} \geq 0$, the model in (1.1) might be rewritten as a random effects model (Laird and Ware, 1982), using standard software like the packages "lme" (Bates et al., 2015) or "nlme" (Pinheiro et al., 2015) inside R (R Core Team, 2014). It is worth noting that these packages utilize the REML principles for estimating the variance components that are central in the random effects models.

Finally, by a slight transformation of the model, the data might be analysed using a method known as two-stage linear regression (2SLS) (Wooldridge, 2012, chap.15). 2SLS is a method well–known in econometrics. The method is easy to implement using the package "systemfit" (Henningsen and Hamann, 2007) inside R.

2 The empirical Bayes machinery

The following section is an excerpt of well–known theory. This theory describes a framework of advantageous methods that were implemented, or might be easily implemented, for the results derived in papers I–III, but also to a large extent for the results of paper V.

2.1 Bayes theorem – standard methods

The cornerstone of Bayesian statistics is the iconic Bayes theorem:

$$p\left(\boldsymbol{\theta}_i \mid \boldsymbol{Y},\, \boldsymbol{\eta}_i\right) = \frac{f\left(\boldsymbol{Y} \mid \boldsymbol{\theta}_i\right) g\left(\boldsymbol{\theta}_i \mid \boldsymbol{\eta}_i\right)}{m\left(\boldsymbol{Y} \mid \boldsymbol{\eta}_i\right)},\;\; i = 1, \ldots, l; \tag{2.2}$$

The notation in (2.2) is based on the notation in Carlin and Louis (2008), and is suitable for explaining the general theory in the next paragraphs. However, it differs slightly from the notation used in paper II.

We assume that there exist a set, $\{M_1, \ldots, M_l\}$, of l different possible models that might explain the data, $\boldsymbol{Y}$, indicated by the sub–indexing $_i$ in (2.2). The posterior distribution of the model parameters, $\boldsymbol{\theta}_i$, conditional on data and hyperparameters, $\boldsymbol{\eta}_i$, is denoted $p(\boldsymbol{\theta}_i \mid \boldsymbol{Y}, \boldsymbol{\eta}_i)$. The likelihood, in our case defined by (1.1), is denoted $f(\boldsymbol{Y} \mid \boldsymbol{\theta}_i)$ and the prior distribution of the parameters is denoted $g(\boldsymbol{\theta}_i \mid \boldsymbol{\eta}_i)$. In common the conditioning on hyperparameters, $\boldsymbol{\eta}_i$, and model indexing is dropped, but this refinement is useful for the present section. However, where conditioning on model is redundant I have omitted this indexing.

The denominator in (2.2), $m(\boldsymbol{Y} \mid \boldsymbol{\eta}_i)$, is known as the model evidence or the marginal likelihood. It might be viewed as a normalizing constant that ensures that the posterior distribution integrates to one over the model parameters, $\boldsymbol{\theta}_i$:

$$m(\boldsymbol{Y} \mid \boldsymbol{\eta}_i) = \int_{\boldsymbol{\theta}} \left[f(\boldsymbol{Y} \mid \boldsymbol{\theta}_i)\, g(\boldsymbol{\theta}_i \mid \boldsymbol{\eta}_i) \right] \delta\boldsymbol{\theta}_i;$$

, thus, assuming that the prior distribution is a proper probability distribution (PDF), the posterior distribution is also a proper PDF for $\boldsymbol{\theta}_i$.

After the advent of computer technology and the possibility to conduct effective data simulations, the popularity of Bayesian statistics increased sharply. The method known as Marcov Chain Monte Carlo sampling enables examination of the posterior distribution, even for extremely complex models, without the need of calculating the model evidence (Gilks et al., 1996; Kass et al., 1998).

A specific class of prior distribution, known as conjugate priors, has the property that their corresponding posterior distributions are known PDF–s of the same family of distributions as the conjugate prior. The term conjugate priors was first used by Schlaifer and Raiffa (1961), but the theory was developed independently by G. A. Barnard and reported by Wetherill (1961), according to David and Edwards (2001).

For models where a conjugate prior distribution exists, a Bayesian analysis of the posterior distribution is simple, as the posterior distribution is a known PDF for which parameters are given by known functions of data, $\boldsymbol{Y}$, and hyperparameters, $\boldsymbol{\eta}$. Often, as is the case of the model described in paper II, the conjugate distributions are compound PDF–s. However, the analysis is still simple, as Monte Carlo simulations for all parameters are achievable without the need to implement Markov Chains. Furthermore, analytical expressions based

on data and hyperparameters for basic properties like (posterior) means, medians and variances for the variable (in the Bayesian sense) in question, $\boldsymbol{\theta}$, will often be achievable.

2.2 Model selection

Another asset by applying a conjugate prior is the possibility to calculate the model evidence. An analytical expression is easily derived by a slight transformation of (2.2), utilizing the fact that $p(\boldsymbol{\theta}_i \mid \boldsymbol{Y}, \boldsymbol{\eta}_i)$, $f(\boldsymbol{Y} \mid \boldsymbol{\theta}_i)$ and $g(\boldsymbol{\theta}_i \mid \boldsymbol{\eta}_i)$, are all known functions.

Since a possible interpretation of the model evidence is "likelihood of data conditional on model and hyperparameters", model evidence might be used for model selection (Kass and Raftery, 1995; MacKay, 1992). The well–known "Bayes factor" for comparing two models is simply the ratio of their model evidences.

Under the assumption that one of the possible models, $\{M_1, \ldots, M_l\}$, is the correct model, a posterior probability for a model being the correct one is easily achievable within the Bayesian regime as:

$$\pi(M_i \mid \boldsymbol{Y}, \boldsymbol{\eta}_i) = \frac{m(\boldsymbol{Y} \mid \boldsymbol{\eta}_i)\gamma(M_i)}{\sum_{i=1}^{l}[m(\boldsymbol{Y} \mid \boldsymbol{\eta}_i)\gamma(M_i)]}; \tag{2.3}$$

, where $\pi(M_i \mid \boldsymbol{Y}, \boldsymbol{\eta}_i)$ and $\gamma(M_i)$ denotes the posterior and prior probabilities that model M_i is the correct model. If one model is to be chosen, the model with the largest posterior probability should be used. Under the natural, though not mandatory, flat prior distribution, i.e. $\gamma(M_i) = 1/l$ for $i = 1, \ldots, l$, the model with largest model evidence would also be the model with largest posterior probability. In paper III this method, i.e. choosing the model with largest model evidence, is used for model selection.

Yet another variant of utilizing Bayesian methods for model selection in ordinary linear regression, i.e. univariate response, is described in George and Foster (2000), where the same model hyperparameters, $\boldsymbol{\eta}$, are applied to all combination of predictor variables, i.e. $\boldsymbol{\eta}_i = \boldsymbol{\eta}_j$ for all $i = 1, \ldots, l,\ \ j = 1, \ldots, l$.

2.3 Model averaging

Selecting a single model ignores the model uncertainty. In turn, this leads to underestimation of uncertainty about the quantities of interest, which for this thesis are the predicted values. A possible solution to deal with the model uncertainty is to apply methods known as "Bayesian model averaging". The natural Bayesian solution for prediction would be to do the prediction using all possible models and then apply a weighted average of these

predictions, as the final prediction. The weights involved would be defined by the posterior probabilities for models as given by (2.3).

Even though this method might be shown to provide optimal prediction ability (Madigan and Raftery, 1994), it will be unsuitable for a lot of practical situations. Raftery et al. (1997) outlines two practical approaches for "Bayesian model averaging" in ordinary linear regression. I have not pursued this method for model averaging in the case with missing data, but assume that the principles from Raftery et al. (1997) should be transmittable to the situation with missing data. The principles of Bayesian model averaging has met with some criticism, and as already mentioned, a frequentist alternative is proposed by Hjort and Claeskens (2003).

2.4 Empirical Bayes

The principles outlined in this section are valid for a more general linear regression model than the model defined in (1.1), see Algorithm a.1 for details. The model defined by (1.1) fits into this general framework.

So far I have considered the hyperparameters, $\boldsymbol{\eta}$, as known. The basic idea of empirical Bayes methods is to set the hyperparameters based on the data. Since the model evidence, $m(\boldsymbol{Y} \mid \boldsymbol{\eta})$, might be viewed as a likelihood function like in (2.3), a common method for setting the hyperparameters is to use the maximum likelihood (ML) estimate, i.e. $\boldsymbol{\eta} = argmax_{\boldsymbol{\eta}}\left[m(\boldsymbol{Y} \mid \boldsymbol{\eta})\right]$. This optimisation is not necessarily solvable by analytical methods. A simple and practical alternative, assuming the dimension of $\boldsymbol{\eta}$ is not too high, is to use a non–linear numerical optimizer for calculating $\boldsymbol{\eta}$. We utilized this method in both paper II and paper III. Another possibility is to use the expectation maximization algorithm for setting hyperparameters such that model evidence is maximized.

The Bayesian estimator for regression parameters for model $_i$ is given by $\tilde{\boldsymbol{\beta}}_i$, which under quadratic loss function is the posterior mean. Further, let a new observation be given by $\{\boldsymbol{y}_N,\ \boldsymbol{x}_N\}$, where $\boldsymbol{y}_N$ denotes a $q \times 1$ vector of new responses, and the $p \times 1$ vector $\boldsymbol{x}_N$ contains the corresponding predictor variables.

The Bayesian prediction, is given by $\tilde{\boldsymbol{y}}_N = \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_N$, where $\tilde{\boldsymbol{\beta}}$ might be estimated by model selection or model averaging. Algorithm a.1 outlines the basic principles for how this prediction is reached by applying the totally data–driven empirical Bayes framework.

In regression, the set of possible models, $\{M_1, \ldots, M_l\}$, denotes the combinations of predictor variables. If we have p possible predictors and want to test "all combinations" we get $l = 2^p$. Thus, if p is large, some kind of screening should be implemented to reduce the

number of possible models, before Algorithm a.1 is applied.

Algorithm a.1 A general algorithm for empirical Bayes prediction in a linear model

Input:

Data:

→ Training set with responses $\boldsymbol{Y}$ $(n_1 \times q)$, and predictors $\boldsymbol{X}$ $(n_1 \times p)$, where subindex $_j$ is used to identify row, i.e. $j = 1, \ldots, n_1$.

→ Predictor variables for a new observation, $\boldsymbol{x}_N$.

Model framework:

→ A set of possible models $\{M_1, \ldots, M_l\}$, where subindex $_i$ is used to identify model, i.e. $i = 1, \ldots, l$.

→ A data structure with independent, identical distributed (*i.i.d.*) observations.

→ Likelihood, $f(\boldsymbol{y_j} \mid \boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{x}_j, M_i)$ such that:

→ $E(\boldsymbol{y_j} \mid M_i) = \boldsymbol{X}\boldsymbol{\beta}_i$

→ $var(\boldsymbol{y_j} \mid M_i) = h(\boldsymbol{\Sigma}_i, \ \boldsymbol{X})$

→ Conjugate prior distribution, $g(\boldsymbol{\beta}_i, \ \boldsymbol{\Sigma}_i \mid \boldsymbol{\eta}_i)$, such that:

→ $m(\boldsymbol{Y} \mid \boldsymbol{\eta}_i)$ has known form.

→ A prior distribution for the model, $\gamma(M_i)$.

Result: A predicted value, $\tilde{\boldsymbol{y}}_N$, for the response in the new observation.

Procedure:

For $i = 1$ to $i = l$

1) Set hyper–parameters based on model and data.

→ *An alternative is the* ML *method, i.e.* $\boldsymbol{\eta}_i = argmax_{\boldsymbol{\eta}}\left[m(\boldsymbol{Y} \mid \boldsymbol{\eta}_i)\right]$

2) Calculate and store the model evidence, $m(\boldsymbol{Y} \mid \boldsymbol{\eta}_i)$

End

3) Estimate $\boldsymbol{\beta}$ by $\tilde{\boldsymbol{\beta}}$.

→ ***Alternative a:*** *By model selection, see section (2.2)*

→ ***Alternative b:*** *By model averaging, see section (2.3)*

4) Predict $\tilde{\boldsymbol{y}}_N = \tilde{\boldsymbol{\beta}}^T \boldsymbol{x}_N$

3 Summaries of papers I–III

3.1 Paper I: Prediction error – missing data

It is worth noting that even though paper I provides strong foreshadowing to paper II, a paper based on Bayesian methods, paper I is based upon ordinary frequentist principles.

In the first paper we made the (unrealistic) assumption of known error covariance struc-

ture, i.e. that $\boldsymbol{\Sigma}$ was known. Matrix algebra showed that the generalized least squares (GLS) estimator (Aitkin, 1935) for $\boldsymbol{\beta}_2$, denoted $\hat{\boldsymbol{\beta}}_{v2}$, known by general theory to be the UMVUE estimator (Amemiya, 1985, chap.6), is:

$$\hat{\boldsymbol{\beta}}_{v2} = \hat{\boldsymbol{\beta}}_{22} + \sigma_{12}/\sigma_{11}\left(\hat{\boldsymbol{\beta}}_{11} - \hat{\boldsymbol{\beta}}_{21}\right); \tag{3.4}$$

, where $\hat{\boldsymbol{\beta}}_{21}$ and $\hat{\boldsymbol{\beta}}_{22}$ are the standard OLS estimates based on the n_2 first observations (i.e. the observations without missing data) for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ respectively, and $\hat{\boldsymbol{\beta}}_{11}$ is the OLS estimate for $\boldsymbol{\beta}_1$ based on the full set of data (all n_1 rows).

A crucial point, which is evident by simple analysis of (3.4), is that when the errors of the two responses are uncorrelated, i.e. $\sigma_{12} = 0$, and/ or no data are missing, i.e. $n_2 = n_1$, then the GLS estimator, $\hat{\boldsymbol{\beta}}_{v2}$, simplifies to the OLS estimator based on the full observations, $\hat{\boldsymbol{\beta}}_{22}$. Consequently, in these particular cases, inclusion of a surrogate variable with extra observations has no effect on the UMVUE estimator for $\boldsymbol{\beta}_2$.

As $\hat{\boldsymbol{\beta}}_{v2}$ was shown to be multivariate normal distributed, we were able to analytically confirm that the gain of including a surrogate variable increased as the absolute value of the correlation, $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$, between error terms increased. Furthermore, the gain increased as the difference between n_1 and n_2 increased, an intuitively correct result.

The next step of paper I was to analyse a biased estimator for $\boldsymbol{\beta}_2$, $\tilde{\boldsymbol{\beta}}_2 = k\hat{\boldsymbol{\beta}}_{v2}$, where $0 < k < 1$, commonly known as the James–Stein estimator (James and Stein, 1961). If k is set properly, the James–Stein estimator is a better estimator than the GLS estimator in the sense of having lower risk function, i.e. $R_{\hat{\boldsymbol{\beta}}_{v2}} > R_{\tilde{\boldsymbol{\beta}}_2}$.

As our focus was prediction error, we evaluated how well a new observation, $\boldsymbol{y}_N$, was expected to be predicted based on its corresponding predictor variable, $\boldsymbol{x}_N$. We denoted the prediction based on the James–Stein estimator $\tilde{\boldsymbol{y}}_N$, thus the risk function of interest was $R_{\tilde{\boldsymbol{y}}_N}$. In the first theorem of the paper, we showed the analytical solution for this risk. This risk is a function of k, n_1, n_2, $\boldsymbol{\Sigma}$, $\boldsymbol{\beta}$, $\boldsymbol{X}_1$ and $\boldsymbol{x}_N$.

The standard model for regression analysis does not include assumptions about the predictor variables, $\boldsymbol{X}_1$ and $\boldsymbol{x}_N$. On the contrary, by using design matrices and/ or factorial variables as predictors, the underlying models of the wide scope of statistical methods known as "variance analysis", might be viewed as linear regression models. However, a common method for constructing the matrix of predictor variables, $\boldsymbol{X}_1$, and responses, $\boldsymbol{Y}_1$, is registration of features connected to a common object of interest.

Motivated by Helland and Almøy (1994) we wanted to evaluate aspects of the prediction precision for the James–Stein estimator conditional on known values for some model

parameters; the error variance $\boldsymbol{\Sigma}$, the population coefficients of determination, R_1^2 and R_2^2, and the intercept terms, $\boldsymbol{\beta}_0$, but unconditional on the predictor variables, i.e. $\boldsymbol{X}_1$ and $\boldsymbol{x}_N$. We made the assumptions that all predictors were independently drawn from a common multivariate normal distribution, and that the missing responses were missing at random. By general theory this enabled us to find analytical results for the expected prediction risk, $E_{\boldsymbol{x}}\left(R_{\tilde{\boldsymbol{y}}_N}\right)$, over all training samples, $\boldsymbol{X}_1$, and new prediction variables, $\boldsymbol{x}_N$. These risks, which are functions of k, n_1, n_2, $\boldsymbol{\Sigma}$, R_1^2, R_2^2 and $\boldsymbol{\beta}_0$, constitutes the core of the second theorem of paper I.

The final step of paper I was to show that simulations using the R package "Simrel" (Sæbø, 2015; Sæbø et al., 2015) showed a pattern consistent with the analytical results. Even if these simulations did not prove the analytical results to be correct, they strongly supported the validity of our theoretical findings.

3.2 Paper II: Empirical Bayes – double shrinkage

Life is not always a matter of holding good cards,
but sometimes, playing a poor hand well.
– Jack London.

Papers I and II share the common basic model and problem with missing data. However, in paper II, the unrealistic assumption of known covariance for the error terms was omitted. Thus the methods outlined in paper II are fully serviceable for a practical situation where the model in (1.1) is suitable.

The connection between certain prior distributions, empirical Bayes methods, and James–Stein inspired estimators, was the theme of a series of papers by Efron and Morris (1971, 1972a,b, 1973, 1975, 1976). In paper II we applied these principles to the problem in question. We showed that the resulting empirical Bayes estimator might be viewed as a double shrinkage estimator, whose prediction precision for a large part of model parameter settings outperformed natural competitors.

Paper II contains three crucial developments; (*i*) we showed that a conjugate prior distribution exists for the missing data situation; (*ii*) we showed that shrinkage of the unbiased estimate for the term σ_{12}/σ_{11}, utilized in the regression parameter estimate in (3.4) is beneficial; and (*iii*) we showed how a parametrization with four free parameters, of which two are connected to the estimator for $\boldsymbol{\beta}_2$, yields a desired double shrinkage effect.

The conjugate prior distribution we used might be viewed as a slight transformation of the matrix–normal–inverse–Wishart distribution, which in turn is a conjugate prior distribu-

tion for a linear regression model with multivariate response variable without missing data (Box and Tiao, 1973). The elements of a Wishart distributed variable might be expressed as a compound distribution composed of two new (independent) Wishart distributions and a multivariate normal distribution (Giri, 2003). We used a parameter transformation of $\boldsymbol{\Sigma}$, which utilized this compound form. Since $\boldsymbol{\Sigma}$ in the present case was a 2×2 matrix, the two "new" Wishart distributions simplified to gamma distributions, as the gamma distribution might be viewed a univariate Wishart. The multivariate normal simplified to a univariate normal distribution.

The transformation in Giri (2003) is valid for Wishart distributed variables of all dimensions, corresponding to missing data problems in the general multivariate situation. Thus, I suspect that most of the results obtained in paper II for the bivariate response, might be possible to expand to the general multivariate response model.

Based on the conjugate prior distribution for the model in (1.1), the closed form solution for the model evidence, denoted $\pi(\boldsymbol{Y})$ in paper II, was derived. Consequently, the totaly data–driven Algorithm a.1 might be applied to practical problems.

If we stick to an unbiased estimator for $\boldsymbol{\beta}_2$, it was shown in paper I that the GLS estimator is given by $\hat{\boldsymbol{\beta}}_{v2}$, see (3.4). If $\boldsymbol{\Sigma}$ is unknown, as is the case for all practical situations, the natural solution is to substitute the term σ_{12}/σ_{11} in (3.4) by an estimate. By applying an argument analogous to the arguments leading to the canonical results obtained by Stein et al. (1956), we showed that in order to minimize the variance of the (unbiased) estimator for $\boldsymbol{\beta}_2$, based on (3.4), we should use a biased (i.e. shrunk towards zero) estimate for σ_{12}/σ_{11}. Unfortunately the ideal level of shrinkage is itself a function of the unknown model parameter $\boldsymbol{\Sigma}$.

As already observed, the GLS– and UMVUE–estimator for $\boldsymbol{\beta}_2$ simplifies to the OLS–estimator based on the observations without missing data, if $\sigma_{12} = 0$. Consequently, if $\sigma_{12} = 0$, an estimator based on (3.4), where σ_{12}/σ_{11} is substituted with any estimator having variance larger than zero, will have larger variance than the OLS/GLS/UMVUE–estimator. Both estimators will be unbiased.

Thus, in situations where the true value for σ_{12} is zero or close to zero, the statistician trying to utilize the GLS inspired estimator in (3.4), is "dealt poor cards". In order to play these cards well, the estimator for σ_{12}/σ_{11} should be shrunk very close to zero. Oppositely, when ρ^2 is large, the statistician trying to utilize the GLS inspired estimator, is "dealt good cards". In order to play them well the shrinkage of the estimate for σ_{12}/σ_{11} should be small, for maximizing the gain of the observations with missing data.

The empirical Bayes hyperparameters we applied to the model were heavily inspired by

the prior proposed by Minka (2000), which is based on the principles of the "Zellner's g–prior" (Zellner, 1986). We showed that if it is possible to tune the two free hyper parameters associated with the estimator for $\boldsymbol{\beta}_2$ properly, then we get a "double shrinkage estimator" of the desired form. The tuning was simply conducted by following the empirical Bayes strategy of Algorithm a.1, i.e. by maximizing the model evidence. As a very indicative explanation for why this strategy works, one might view our prior distribution as the support for the hypothesis that $\sigma_{12}/\sigma_{11} = 0$. If the support of this hypothesis, conditional on data, is high, the shrinkage of the estimate for σ_{12}/σ_{11} towards zero is heavy. Oppositely if the support is low, so is the shrinkage.

When working with paper II, we put considerable effort into finding a prior distribution that would result in a posterior distribution for σ_{12}/σ_{11}, utilizing the part of data where one of the responses was missing. We did not succeed. However, I think some of the results we achieved, briefly summed up in section 4, have some general interest, and indicate that there might exist prior distributions leading to additional reduction in the expected prediction error compared to the prior we present in paper II.

The final step of paper II was to use the R-package "Simrel" to simulate data where the true population parameters were changed systematically. The simulations confirmed the general results obtained in paper I, as the gain of substituting the OLS–estimator with the empirical Bayes estimator increased when: (i) ρ^2 increased, (ii) the number of missing data increased, (iii) the coefficient of determination associated with the response without missing data, R_1^2, increased, (iv) the coefficient of determination associated with the response with missing data, R_2^2, decreased. Furthermore, it was shown that tuning the free hyperparameters by maximizing model evidence seems to be a suitable method.

Another aspect that required a considerable amount of extra work, unfortunately without a result, was the search for an analytic solution for setting the hyperparameters, $\boldsymbol{\eta}$. I suspect that such solutions are achievable, for instance using the features of the hypergeometric function, ${}_2F_1$, see Abramowitz and Stegun (1964, chap.15). If so, some of the results we had to evaluate by simulation might be possible to evaluate by direct analytical analysis.

3.3 Paper III: Lean meat percentage in pork

As already pointed out, the practical problem addressed in paper III initiated the analysis described in paper I and paper II. Hence, paper III might be viewed as a practical application for the empirical Bayes method described in paper II. The short conclusion is: The empirical Bayes method performed well on a set of real data, where measurements from the optical

probe Hennessy Grading Probe 7 (HGP7) combined with some other predictor variables, were used to predict lean meat percentage (LMP) in Norwegian pig carcasses.

A model using four predictor variables, three measurements from HGP7, and gender, was simple and provided a high prediction precision for LMP, well inside EU standards. For practical simplicity gender might be omitted without severe loss of prediction precision.

We also showed that the number of pigs manually dissected could be substantially reduced without severe loss of prediction precision, provided a sufficiently large number of CT scanned carcasses. This aspect was not addressed in paper I or II. A key lesson was that if one possesses some prior knowledge about the covariance matrix $\boldsymbol{\Sigma}$, the coefficients of determination, R_1^2 and R_2^2, the cost associated with sampling the responses, and the gain of reducing prediction error, this prior knowledge might be used for optimizing the sampling design of the experiment.

4 An alternative estimator for σ_{12}/σ_{11}

This section treats an estimator for the proportion σ_{12}/σ_{11} as defined in paper I and II. The results presented in this section are not required to understand the entirety of the rest of the thesis. However, the results indicate that there may exist prior distributions that are potentially "better" than the prior proposed in paper II. In paper I and II we assumed that $\boldsymbol{Z}$, i.e. the last $p-1$ columns of $\boldsymbol{X}$, were centred and normally distributed. These assumptions are not necessary for the results derived in this section, where I introduce:

$$q_4 = \boldsymbol{y}_{-\delta 1}^T \left(\boldsymbol{I}_{(n_1 - n_2)} - \boldsymbol{H}_{-\delta} \right) \boldsymbol{y}_{-\delta 1};$$

, where the subindex $_{-\delta}$ indicates the rows containing missing data. $\boldsymbol{Q}$ is defined in paper I and II.

Proposition 1 - An alternative estimator for σ_{12}/σ_{11}
Under the assumption that ρ^2 follows a proper prior PDF *with known expected value, $E(\rho^2)$, I propose to use the following estimator for σ_{12}/σ_{11}:*

$$\begin{aligned} \widehat{\sigma_{12}/\sigma_{11}} =& a\, q_{12}/q_{11} + (1-a)\left[(n_1 - n_2 - p - 2)\, q_{12}\right]/\left[(n_2 - p)\, q_4\right]; \\ & \text{where:} \\ a =& \begin{cases} 1, & \text{if } n_1 \leq n_2 + p + 4 \\ \frac{(n_2-p-2)\left[1+(n_1-2p-3)E\left(\rho^2\right)\right]}{(n_2-p-2)[1+(n_1-2p-3)E(\rho^2)]+(n_1-n_2-p-4)[1-E(\rho^2)]}, & \text{else;} \end{cases} \end{aligned} \tag{4.5}$$

Lemma 1

The estimator proposed in (4.5) is unbiased for all $0 < a < 1$*. The value for* a *as given by (4.5) minimizes the variance of* $\widehat{\sigma_{12}/\sigma_{11}}$ *as given in (4.5).*

Proof

Since q_4 is based solely on the observations containing missing data, $1/q_4$ and $\boldsymbol{Q}$ are independent variables, whose distributions are given by:

$$\begin{aligned} 1/q_4 \sim & IG\left[(n_1 - n_2 - p)/2,\ \ 1/(2\sigma_{11})\right]; \\ \boldsymbol{Q} \sim & W\left[\boldsymbol{\Sigma},\ \ (n_2 - p)\right]; \end{aligned}$$

Due to the Wishart distribution of $\boldsymbol{Q}$ we have (Giri, 2003):

$$\begin{aligned} \left(q_{22} - q_{12}^2/q_{11}\right) \sim & Ga\left\{(n_1 - n_2 - p - 1)/2,\ \ 1/\left[2\left(\sigma_{22} - \sigma_{12}^2/\sigma_{11}\right)\right]\right\}; \\ \left(q_{12}/q_{11},\ \ q_{11}\right) \sim & NGa\left[\sigma_{12}/\sigma_{11},\ \ 1/\left(\sigma_{22} - \sigma_{12}^2/\sigma_{11}\right),\ \ (n_2 - p)/2,\ \ 1/(2\sigma_{11})\right]; \end{aligned}$$

, where $(q_{22} - q_{12}^2/q_{11})$ and $(q_{12}/q_{11},\ \ q_{11})$ are independent. To get the expression for the covariance between q_{12}/q_{11} and q_{12}/q_4 I used that:

$$E\left[q_{12}^2/(q_{11}q_4)\right] = -\left[E\left(q_{22} - q_{12}^2/q_{11}\right) - E\left(q_{22}\right)\right] E\left(1/q_4\right);$$

Then I get the following expressions for expected values, variances and covariance for the variables q_{12}/q_{11} and q_{12}/q_4:

$$\begin{aligned} E\left(q_{12}/q_{11}\right) =& \sigma_{12}/\sigma_{11}; \\ var\left(q_{12}/q_{11}\right) =& \left(\sigma_{22}/\sigma_{11}\right)\left(1 - \rho^2\right)\left[1/(n_2 - p - 2)\right]; \\ E\left(q_{12}/q_4\right) =& \left(\sigma_{12}/\sigma_{11}\right)\left[(n_2 - p)/(n_1 - n_2 - p - 2)\right]; \\ var\left(q_{12}/q_4\right) =& \left(\sigma_{22}/\sigma_{11}\right)\left(\left\{(n_2 - p)/[(n_1 - n_2 - p - 2)\left(n_1 - n_2 - p - 4\right)]\right\} + \right. \\ & \left.\left\{\left[(n_2 - p)\left(n_1 + n_2 - 3p - 2\right)\right]/\left[\left(n_1 - n_2 - p - 2\right)^2\left(n_1 - n_2 - p - 4\right)\right]\right\}\rho^2\right); \\ cov\left(q_{12}/q_{11},\ \ q_{12}/q_4\right) =& \left(\sigma_{22}/\sigma_{11}\right)\left(1 - \rho^2\right)\left[1/(n_1 - n_2 - p - 2)\right]; \end{aligned}$$

This leads to:

$$\begin{aligned} E\left(\widehat{\sigma_{12}/\sigma_{11}}\right) =& \sigma_{22}/\sigma_{11}; \\ var\left(\widehat{\sigma_{12}/\sigma_{11}}\right) =& \left(\sigma_{22}/\sigma_{11}\right)\left[f_1\left(a\right) - f_2\left(a\right)\rho^2\right]; \end{aligned}$$

, where:

$$f_1(a) = \left\{ \left[2(n_2 - p - 2)a - (n_2 - p - 4)a^2\right] / [(n_2 - p)(n_2 - p - 2)] + \left[(1-a)^2 (n_1 - n_2 - p - 2)\right] / [(n_1 - n_2 - p - 4)(n_2 - p)] \right\};$$

$$f_2(a) = \left\{ \left(2(n_2 - p - 2)a - (n_2 - p - 4)a^2\right) / [(n_2 - p)(n_2 - p - 2)] - \left[(1-a)^2 (n_1 + n_2 - 3p - 2)\right] / [(n_1 - n_2 - p - 4)(n_2 - p)] \right\};$$

By viewing ρ^2 as a random variable with expectation $E(\rho^2)$, a Bayesian approach, and applying the law of total variance (Eve's law) the variance of $\widehat{\sigma_{12}/\sigma_{11}}$ is:

$$var\left(\widehat{\sigma_{12}/\sigma_{11}}\right) = (\sigma_{22}/\sigma_{11}) \left[f_1(a) - f_2(a) E\left(\rho^2\right)\right];$$

, which is minimized for a as given in (4.5). One should note that:

$$var_{\rho^2}\left[E\left(\widehat{\sigma_{12}/\sigma}11\right)\right] = var_{\rho^2}(\sigma_{12}/\sigma_{11}) = 0;$$

References

A. P. Dempster, N. M. Laird, D. B. R. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

Abramowitz, M. and Stegun, I. A., editors (1964). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*. U.S. Government Printing Office, Washington, D.C.

Aitkin, A. (1935). On least squares and linear combination of observations. In *Proceedings of the Royal Society of Edinburgh*, volume 55, pages 42–48.

Amemiya, T. (1985). *Advanced Econometrics*. Harvard University Press.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed–Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Box, G. and Tiao, G. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley series in behavioral science: quantitative methods. Addison-Wesley Pub. Co.

Carlin, B. P. and Louis, T. A. (2008). *Bayesian Methods for Data Analysis*. CRC Press.

Claeskens, G. and Hjort, N. L. (2003). The Focused Information Criterion. *J. Amer. Statist. Assoc.*, 98(464):900–916.

Daumas, G. and Monziols, M. (2011). Comparison between computed tomography and dissection for calibrating pig classification methods. In *57th International Congress of Meat Science and Technology (ICoMST 2011)*, volume 1, pages 296–299. Ghent-Belgium, 7–12 August 2011.

David, H. A. and Edwards, A. W. F. (2001). First (?) Occurrence of Common Terms in Statistics and Probability. In David, H. A. and Edwards, A. W. F., editors, *Annotated Readings in the History of Statistics*, chapter Appendix B, pages 219–228. Springer Science & Business Media.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford,UK.

Efron, B. and Morris, C. (1971). Limiting the Risk of Bayes and Empirical Bayes Estimators-Part I: The Bayes Case. *J. Amer. Statist. Assoc.*, 66(336):807–815.

Efron, B. and Morris, C. (1972a). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika*, 59(2):335–347.

Efron, B. and Morris, C. (1972b). Limiting the Risk of Bayes and Empirical Bayes Estimators-Part II: The Empirical Bayes Case. *J. Amer. Statist. Assoc.*, 67(337):130–139.

Efron, B. and Morris, C. (1973). Stein's Estimation Rule and its Competitors-An Empirical Bayes Approach. *J. Amer. Statist. Assoc.*, 68(341):117–130.

Efron, B. and Morris, C. (1975). Data Analysis Using Stein's Estimator and its Generalizations. *J. Amer. Statist. Assoc.*, 70(350):311–319.

Efron, B. and Morris, C. (1976). Families of Minimax Estimators of the Mean of a Multivariate Normal Distribution. *Ann. Statist.*, 4:11–21.

George, E. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). Introducing Markov Chain Monte Carlo. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, chapter 1, pages 1–19. London: Chapman and Hall.

Giri, N. C. (2003). *Multivariate Statistical Analysis: Revised and Expanded*, volume 171. CRC Press.

Harville, D. A. (1977). Maximum Likelihood Approaches to Variance Component Estimation and to Related Problems. *J. Amer. Statist. Assoc.*, 72(358):320–338.

Helland, I. S. and Almøy, T. (1994). Comparison of Prediction Methods when Only a Few Components are Relevant. *J. Amer. Statist. Assoc.*, 89(426):583–591.

Henningsen, A. and Hamann, J. D. (2007). systemfit: A Package for Estimating Systems of Simultaneous Equations in R. *Journal of Statistical Software*, 23(4):1–40.

Hjort, N. L. and Claeskens, G. (2003). Frequentist Model Average Estimators. *J. Amer. Statist. Assoc.*, 98(464):879–899.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379.

Kass, R. E., Carlin, B. P., Gelman, A., and Neal, R. M. (1998). Markov Chain Monte Carlo in Practice: A Roundtable Discussion. *The American Statistician*, 52(2):93–100.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, 90(430):773–795.

Laird, N. M. and Ware, J. H. (1982). Random–Effects Models for Longitudinal Data. *Biometrics*, 38(4):963–974.

MacKay, D. J. (1992). Bayesian interpolation. *Neural computation*, 4(3):415–447.

Madigan, D. and Raftery, A. E. (1994). Model Selection and Accounting for Model Uncertainty in Graphical Models Using Occam's Window. *J. Amer. Statist. Assoc.*, 89(428):1535–1546.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. Academic press.

Minka, T. (2000). Bayesian linear regression. Technical report, Microsoft Research Cambridge.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2015). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-121.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian Model Averaging for Linear Regression Models. *J. Amer. Statist. Assoc.*, 92(437):179–191.

Sæbø, S. (2015). *simrel: Linear Model Data Simulation and Design of Computer Experiments*. R package version 1.1-0.

Sæbø, S., Almøy, T., and Helland, I. S. (2015). simrel-A versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors. *Chemometrics and Intelligent Laboratory Systems*.

Savage, L. J. (1976). On rereading RA Fisher. *The Annals of Statistics*, pages 441–500.

Schlaifer, R. and Raiffa, H. (1961). *Applied Statistical Decision Theory*. Boston: Clinton Press, Inc.

Srivastava, J. (1965). A multivariate extension of the Gauss–Markov theorem. *Annals of the Institute of Statistical Mathematics*, 17(1):63–66.

Stein, C. et al. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In *Proceedings of the Third Berkeley symposium on mathematical statistics and probability*, volume 1, pages 197–206.

Upton, G. and Cook, I. (2014). *A Dictionary of Statistics 3e*. Oxford university press.

Vester-Christensen, M., Erbou, S. G., Hansen, M. F., Olsen, E. V., Christensen, L. B., Hviid, M., Ersbøll, B. K., and Larsen, R. (2009). Virtual dissection of pig carcasses. *Meat science*, 81(4):699–704.

Walstra, P. and Merkus, G. (1995). *Procedure for the assessment of lean meat percentage as a consequence of the new EU reference dissection method in pig carcass classification*. DLO Research Institute of Animal Science and Health (ID-DLO), Zeist, The Netherlands.

Wetherill, G. (1961). Bayesian Sequential Analysis. *Biometrika*, 48(3–4):281–292.

Wooldridge, J. M. (2012). *Introductory Econometrics: a Modern Approach*. Cengage Learning.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243.

CT in Porcine

1 Background

The first computet tomography (CT) scanner was developed by Sir Godfrey Hounsfield, and the first patient brain–scan was done on October 1st 1971, according to Beckmann (2006). Hounsfield received the Nobel prize for physiology or medicine together with Allan McLeod Cormack in 1979, for his part in developing the diagnostic technique of CT. The data in CT scans might be viewed as a 3D volume consisting of voxels, where voxel is the basic 3D unit. An intensity measured in the Hounsfield scale (HU) is aligned to each voxel.

The main application for CT is of course human medicine, where huge resources are allocated to the research field broadly known as "medical image analysis", a field covering more techniques than CT. However, at least compared to the age of the technology, CT has a long history within the meat industry, where the first paper utilizing CT on pigs was published in 1981 (Skjervold et al., 1981), only two years after Hounsfield received the Nobel prize.

CT has been used on living livestock animals (*in vivo*) or on carcasses (*post mortem*), for a variety of species, including pigs, sheep, hens/ broilers and calves, see Scholz et al. (2015) for an overview. In general CT has been a very valuable, and accurate, tool for measuring body composition both *in vivo* and *post mortem*.

The breeding company Topigs Norsvin uses CT to measure body composition and monitor orthopaedic disorders in 3 500 breeding boars annually as an integrated part of their testing system. As we had access to these data; we had more than enough data at our disposal to construct our atlas. Without having the full overview, I will also hypothesise that few, if any, studies of humans has access to an equally large amount of empirical CT data.

2 Atlas segmentation

A natural next step in order to exploit information from CT scans in animal breeding and meat sciences, is to find methods for segmenting out and identify specific parts or organs that are of special interest. There are a variety of methods that might be used for such segmentation, of which many requires a substantial proportion of manual assistance.

A method which has a number of advantageous properties is known as "atlas segmentation". The method is well known in human medicine and has been applied to numerous human organs like the brain, heart, liver etc. I will not pretend to have anything even close to a full overview of the whole discipline. However, one should note that the main scope

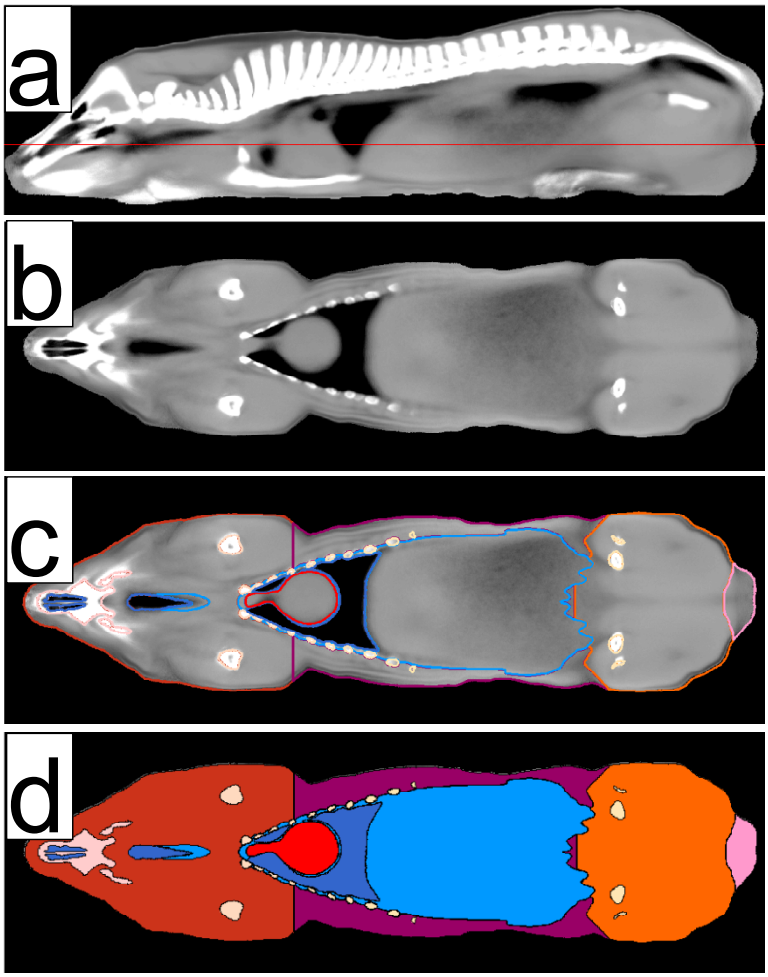


Figure 2.1: (a) The center of the intensity atlas viewed perpendicular to the sagittal plane. Red line show the position of the slice shown in panels b–d. (b) Intensities ("HU–s") for a slice viewed perpendicular to the coronal plane. (c) Boarders for labels in the same slice as shown in panel b. (d) Labels in the same slice as shown in panels b and c. The following labels are shown: Shoulder (red), belly (violet), ham (orange), lungs (dark blue), "guts" (light blue), heart (red), bones (different shades of gray), and testicles (pink).

for human body atlases is single organs, or parts of organs, not the full body. Indeed, a crucial part of the software we utilized in order to fit the non–rigid, cubic B–spline based, transformation of individual pigs to the atlas was originally used for fitting transformations of the human mandible (Kroon, 2011a,b).

Atlas segmentation has actually been applied to pig carcasses in Denmark (Hansen, 2010; Olsen et al., 2012; Vester-Christensen, 2008), where the atlas construction was done fairly easily by simply using a close to average carcass as the atlas. For various reasons, the practical use of the Danish pig carcass atlas, named "Saerimner", has stagnated since its origin in 2010.

As we wanted to construct an atlas for pigs *in vivo*, symmetric over the sagittal plane except for the internal organs, we started the work from scratch. The methods we used,

described in paper V, were to some extent novel. Our main inspiration and guidance came from studies on micro–CT scans of mice, see Segars et al. (2004) for making and labelling the atlas, and Baiker et al. (2010); Li et al. (2008) for automatic non–rigid registration of individual bodies to an atlas. Even if mice are animals, they are not livestock animals. Thus we exploited information from studies primarily aimed at human medicine (laboratory mice), for agricultural purposes.

The labelled atlas that we constructed is deterministic, i.e. every voxel is aligned to a label with probability 0 or 1 (Fig. 2.1). Another type of atlases is a so called probabilistic atlas, where every voxel is aligned to a label with a probability between 0 and 1. For the pig atlas the gain of applying a probabilistic atlas seems limited at the moment, as the labelled parts are large, and the segmentation rough. However, the intensity atlas allows for a substantially higher level of detail in the labelled atlas. The usefulness of a probabilistic atlas will increase as the level of detail increases.

Probabilistic atlases might be constructed in different ways. The simplest method is to manually segment, i.e. define the labels of, a (sub)sample of the individuals used for the atlas construction. After registration of these individuals to the atlas, the probability of a random atlas voxel being aligned to a label might be estimated by the corresponding Monte Carlo estimate. Van Leemput (2009) describes a more sophisticated Bayesian inspired method for constructing a probabilistic atlas. This method also relies on a subset of manually segmented atlases as input.

A probabilistic atlas might also be constructed using an intensity atlas as input. Such constructions are not a topic of this thesis, but will often involve variations of methods based on Markov random fields, or related principles. To the best of my knowledge, I assume that the intensity atlas for *in vivo* pigs would constitute a solid basis for such methods. It might also be possible to utilise other features than *average* intensity for each voxel. Voxel–wise features for intensities like *variance*, *quantiles* etc. would be easy to calculate. Such features might be helpful in the process of constructing both deterministic and probabilistic atlases.

3 Summaries of papers IV and V

3.1 Paper IV: Skeleton segmentation

An automatic method for segmenting and identifying the major bones in CT scans of pigs, *in vivo* or *post mortem*, is described in paper IV. The underlying motivation for developing this method was that the skeleton constitutes the basic framework for constructing a full

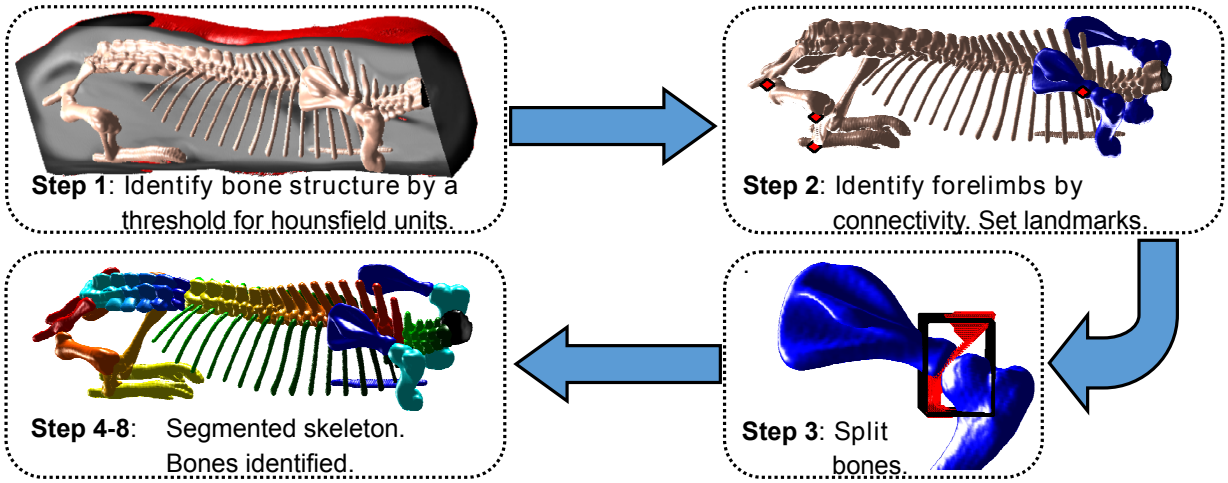


Figure 3.2: Reproduction of the Graphical Abstract available for the online version of paper IV. The figure illustrates the work flow used for bone segmentation and identification.

body atlas, as described in paper V.

The work flow we applied is described through the following enumerated list. The steps in figure 3.2 correspond to this numbering.

1. The CT image was uploaded. The skeletons were segmented out by applying a threshold value (180HU).

2. Skeletons were segmented into three identified major parts, two forelimbs and the "central skeleton" by identification of connected objects.

3. The two forelimbs were segmented into three parts.

4. The two hindlimbs were segmented from the rest of the "central skeleton".

5. The two hindlimbs were segmented into 4 smaller parts.

6. The rest of the "central skeleton" were segmented into 7 main parts.

7. The individual ribs were segmented out and identified.

8. The individual neck vertebras, thoracic vertebras, and lower back vertebras were segmented out and identified.

For segmentation of the individual bones, we had to identify a region of interest (ROI), which embraced the physical interface between two bones. Then we used a 3D expansion of Dijkstra's algorithm in order to segment the two bones. The algorithm ensured that the two bones were separated by a continuous, connected surface.

The methods for identifying the ROI–s, were *ad.hoc.*–methods based on *prior* anatomical knowledge. The *prior* knowledge we utilized to make the methods work, might always be a target for improvement. The work carried out into preparing these algorithms was not principally innovative. However, ascertaining how to combine different, well known image analysis techniques like thresholding for intensities and Euclidian distances, using gradients, etc. was both cumbersome and involved a significant proportion of innovation. Not least, writing the computer code for the algorithms was time–consuming work.

3.2 Paper V: Atlas construction

Paper V describes how we constructed an atlas for *in vivo* pigs. The work described in paper IV constituted an important basis for paper V. Each step in the construction process was characterized by being simple, i.e. they were based on basic methods inside the large discipline broadly known as image analysis.

The procedure was applied to a substantial number of individual pigs, 386 in all. All individuals were scanned with high level of detail, i.e. slice–thickness at 1.25 mm, giving approximately 3×10^8 voxels that had to be handled some way or another for each individual. Thus, we faced problems broadly known as big–data challenges; high computational time, lack of memory in computers etc. Consequently, even if the method in principle was simple, the many subsequent steps, each depending on a successful preceding step, mostly involving large sets of data, made the full process challenging to implement. The steps for constructing the intensity atlases were as follows, where the steps a–d correspond to panels a–d in Figure 1 and steps e–h correspond to panels a–d in Figure 2 of paper V:

(a) We segmented and identified all major bones in the 386 pigs as described in paper IV.

(b) We calculated image moment invariants for all bones. These included centre of mass (COM), volume and spatial orientation.

(c) For all individual bones (femur, pelvis etc.), an average bone was constructed by an affine transformation to a common basis, using the moment image invariants for all 386 pigs. The averages for the image moment invariants were calculated.

(d) A grid of landmarks was set on the average bones.

(e) The landmarks were transformed back to all individual pigs by applying the "inverse transformation" based on the moment's invariants.

(f) The "average" landmarks were transformed into the atlas space, constituting the basis of the atlas.

(g) A B-spline based non–rigid transformations, based on the corresponding landmarks in the individual pigs and atlas space, was fitted to transform objects to atlas space. These transformations were applied to the surface–points (skin) of the individual pigs. An average surface (skin) based on these transformed points was constructed, and new sets of corresponding landmarks were set based on Euclidian shortest distances.

(h) New transformations were fitted based on landmarks in skeletons and surface (skin). These transformations were applied to the full set of voxels in all pigs. The result was the intensity atlas. The intensities of each voxel in the intensity atlas were calculated as the average HU from voxels in individual pigs transformed to atlas space.

A labelled atlas, identifying the four major commercial cuts; "shoulder", "loin", "belly" and "ham", the "guts", lungs, heart and the individual bones, was constructed by manual segmentation of the intensity atlas. We demonstrated how the labelled atlas might be used for atlas segmentation. Finally we showed that atlas segmentation of carcasses made a significant contribution to knowledge about the relative sizes of the main commercial cuts in a test set of 52 pig carcasses, when compared with results from manual butchering.

4 Further work – implementation of empirical Bayes

You will never reach your destination
if you stop and throw stones at every dog that barks.
– Winston Churchill.

In this section I will outline an alternative strategy for how the empirical Bayes machinery might be utilized in order to fit the 3D image transformations described in paper V. I want to emphasize that this does not imply any criticism of the method described by Kroon (2011b), which worked flawlessly.

Let the volume of a random pig be defined by the $N_x \times 3$ matrix $\boldsymbol{X}$, where each row of $\boldsymbol{X}$ defines the (Cartesian) coordinates of one voxel. The sum of all N_x voxels constitutes the full volume. The ultimate goal of the registration of a random pig to the atlas is to map $\boldsymbol{X}$ to its corresponding coordinates in the atlas, given by the $N_x \times 3$ matrix $\boldsymbol{Y}$. Then we might apply the model:

$$\boldsymbol{Y} = \boldsymbol{\Phi}_x\boldsymbol{\beta} + \boldsymbol{E},\ \ \boldsymbol{e}_i \overset{i.i.d}{\sim} N_3\left(\boldsymbol{0},\ \ \boldsymbol{\Sigma}\right),\ \ i = 1, \ldots, N_x; \tag{4.1}$$

, where $\boldsymbol{\Phi}_x$ denotes the $N_x \times p$ matrix of basis functions, i.e. a matrix that is a function of $\boldsymbol{X}$. As usual the regression parameters are given by the $p \times 3$ matrix $\boldsymbol{\beta}$, and the error covariance term is given by the 3×3 matrix $\boldsymbol{\Sigma}$. Finally the $N_x \times 3$ matrix $\boldsymbol{E}$ denotes the errors, where row i is denoted $\boldsymbol{e}_i$.

One should note that the underlying idea for the model in (4.1) is that every voxel, represented by its spatial position, $\boldsymbol{x}_i$, $i = 1, \ldots, N_x$, inside the volume of a random pig, has a corresponding "correct" spatial position, $\boldsymbol{y}_i$, inside the atlas. In papers I and II, we evaluated the expected prediction error, $R_{\tilde{y}}$. Due to the quadratic loss function and the Pythagorean quadruple, a natural interpretation of the prediction risk function, $R_{\tilde{y}}$, is "the expected square of the spatial distance between the transformed point and its correct position in the atlas". Under this setup, we see that the landmark based transformation might be seen as a large prediction challenge, where the "atlas position" of approximately 10^8 voxels in each pig is to be predicted by estimates based on approximately 2000 landmarks.

Even though we invested a substantial amount of work in order to gain the results described in papers IV and V, one might, slightly exaggerating, view the results as a way of constructing a multivariate response, i.e. the coordinates of landmarks in the atlas, given by the $n \times 3$ matrix $\boldsymbol{Y}_L$, and 386 different inputs for possible predictors, i.e. the 386 sets of landmark coordinates in the individual pigs, given by the $n \times 3$ matrices denoted $\boldsymbol{X}_L$ for each individual. In paper V we used these data as input into what might be viewed as a "gray box", i.e. the functions from Kroon (2011a,b). Even though we were well aware of the basic principles of the "gray box", we did not audit all computer code.

This "gray box" completed three sequential operations; (*i*) model selection, i.e. optimizing the knot sequens of the B-spline basic functions, (*ii*) estimation of the regression parameters, and (*iii*) applied the transformation to the full volume, i.e. prediction. These operations would fit like a glove for the empirical Bayes machinery as described by Algorithm a.1.

A conjugate prior distribution for the model defined by (4.1) is the matrix–normal–inverse–Wishard distribution (Box and Tiao, 1973), i.e. a compound distribution where the inverse of $\boldsymbol{\Sigma}$ is Wishard distributed and $\boldsymbol{\beta}$, conditional on $\boldsymbol{\Sigma}$ is matrix–normal distributed with $\boldsymbol{\Sigma}$ as one of the scale parameters.

Without going into detail, if the prior expectation for $\boldsymbol{\beta}$ is set to $\boldsymbol{\beta}_0$, then the resulting posterior estimate for $\boldsymbol{\beta}$ will "shrink" the OLS estimator towards $\boldsymbol{\beta}_0$, where the degree of

shrinkage depends on the prior variance for the estimator. A common choice is to set $\boldsymbol{\beta}_0 = \mathbf{0}$, like we did in paper II and III. The prior used by Minka (2000), inspired by the Zellner's g-prior (Zellner, 1986), corresponds to the prior we used in paper II and III, where the second scale matrix is set proportional to $\boldsymbol{X}^T\boldsymbol{X}$ (or $\boldsymbol{\Phi}_x^T\boldsymbol{\Phi}_x$). Another very common prior is to set the second scale matrix proportional to identity. The frequentist counterpart to this prior, when $\boldsymbol{\beta}_0 = \mathbf{0}$, is commonly known as ridge regression.

Since $\boldsymbol{\Phi}_x$ is based on basic functions, in our case cubic B-Splines, a choice that does not affect the following arguments, the second dimension of $\boldsymbol{\Phi}_x$ might in principle be increased to infinity. As the landmark based methods I have described depends on the least squares solution to be computable, the second dimension of $\boldsymbol{\Phi}_x$ can not be larger than the number of landmarks. However, in line with the major theme of this thesis, the second dimension of $\boldsymbol{\Phi}_x$ should be controlled carefully in order to increase the prediction precision. In a highly cited article, MacKay (1992) evaluates how to optimize basic functions like $\boldsymbol{\Phi}_x$ for linear regression, utilizing the principles within the empirical Bayes machinery. Van Leemput (2009) advocates for applying these methods, within the discipline of image analysis for a problem involving construction of an atlas for the human brain.

A huge challenge for the method we have used for the pigs is the uneven dispersal of skeleton based landmarks within the volume. The density of landmarks is especially sparse in and around the belly, but also in parts of the ham. Thus, the transformation applied to points in these areas might be seen as severe extrapolation. A natural way to deal with this uneven dispersal is some kind of weighted estimator. Such estimators should be fairly easy to implement within the principles outlined in this thesis. Typically landmarks in "dense" areas will be assigned small weights and landmarks in "scarce" areas will be assigned large weights. The weights might also reflect the quality of the landmark, i.e. some landmarks might be harder to identify, and thus should be assigned lower weight.

If there exist landmarks that have known correspondence for only one or two of the (cartesian) dimensions of the atlas, the resulting regression challenge will become a missing data problem analogues to the problem analyzed in paper I–III.

As a concluding remark to this chapter on image registration I have an admission to make: I have left out a major part of state–of–the–art image registration principles by restricting all my work to landmark based registration. I have made no attempt to utilize the main data source, i.e. the HU intensities, in the process of fitting the transformations, after the landmarks are identified. I am also (awkwardly) aware, as was briefly commented in paper V, that iterative methods exist, often based on the Gauss–Newton algorithm (Gill and Murray, 1978), that utilize the intensities for fitting the final transformations.

These methods introduce a new loss function, based on a similarity measure between intensities in the individual pigs and the intensity atlas (references and template). Consequently, the nice link between the landmark based methods and linear regression methods that constitutes the main topic of this thesis is broken. However, the atlas and transformations we have fitted constitute a natural starting point for applying Gauss–Newton inspired methods. These methods might be implemented for the full body volume, but as a start, it is natural to implement them for restricted volumes, such as different organs that might be of interest.

References

Baiker, M., Milles, J., Dijkstra, J., Henning, T. D., Weber, A. W., Que, I., Kaijzel, E. L., Löwik, C. W., Reiber, J. H., and Lelieveldt, B. P. (2010). Atlas-based whole-body segmentation of mice from low-contrast Micro-CT data. *Medical Image Analysis*, 14(6):723–737.

Beckmann, E. C. (2006). CT scanning the early days. *The British journal of radiology*.

Box, G. and Tiao, G. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley series in behavioral science: quantitative methods. Addison-Wesley Pub. Co.

Gill, P. E. and Murray, W. (1978). Algorithms for the Solution of the Nonlinear Least-Squares Problem. *SIAM Journal on Numerical Analysis*, 15(5):977–992.

Hansen, M. F. (2010). *The Virtual Knife*. PhD thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark.

Kroon, D.-J. (2011a). B-spline Grid, Image and Point based Registration. MATLAB Central File Exchange. (c) 2009; Dirk–Jan Kroon. Retrieved June 02, 2015.

Kroon, D.-J. (2011b). *Segmentation of the Mandibular Canal in Cone-beam CT Data*. PhD thesis, University of Twente, Enschede, The Netherlands.

Li, X., Yankeelov, T. E., Peterson, T. E., Gore, J. C., and Dawant, B. M. (2008). Automatic nonrigid registration of whole body CT mice images. *Medical physics*, 35(4):1507–1520.

MacKay, D. J. (1992). Bayesian interpolation. *Neural computation*, 4(3):415–447.

Minka, T. (2000). Bayesian linear regression. Technical report, Microsoft Research Cambridge.

Olsen, E., Kjærsgaard, N., and Hviid, M. (2012). Virtual products can be used for optimisation. *Fleischwirtschaft international: Journal for meat production and meat processing*, (6):16–27.

Scholz, A., Bünger, L., Kongsro, J., Baulain, U., and Mitchell, A. (2015). Non–invasive methods for the determination of body and carcass composition in livestock: dual-energy X-ray absorptiometry, computed tomography, magnetic resonance imaging and ultrasound: invited review. *Animal*, pages 1–15.

Segars, W. P., Tsui, B. M., Frey, E. C., Johnson, G. A., and Berr, S. S. (2004). Development of a 4-D Digital Mouse Phantom for Molecular Imaging Research. *Molecular Imaging & Biology*, 6(3):149–159.

Skjervold, H., Grønseth, K., Vangen, O., and Evensen, A. (1981). In vivo estimation of body composition by computerized tomography. *Zeitschrift für Tierzüchtung und Züchtungsbiologie*, 98(1-4):77–79.

Van Leemput, K. (2009). Encoding Probabilistic Brain Atlases Using Bayesian Inference. *Medical Imaging, IEEE Transactions on*, 28(6):822–837.

Vester-Christensen, M. (2008). *Image Registration and Optimization in the Virtual Slaughterhouse*. PhD thesis, Technical University of Denmark, DTU, DK-2800 Kgs. Lyngby, Denmark.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243.

Paper I

Theoretical evaluation of prediction error in linear regression with a bivariate response variable containing missing data.

Lars Erik Gangsei[a,b], Trygve Almøy[b] & Solve Sæbø[b]

Abstract: Methods for linear regression with multivariate response variables are well described in statistical literature. In this study we conduct a theoretical evaluation of the expected squared prediction error in bivariate linear regression where one of the response variables contains missing data. We make the assumption of known covariance structure for the error terms. On this basis, we evaluate three well-known estimators; standard ordinary least squares, generalized least squares and a James-Stein inspired estimator. Theoretical risk functions are worked out for all three estimators to evaluate under which circumstances it is advantageous to take the error covariance structure into account.

Keywords: bivariate linear regression; James–Stein estimator; missing data; prediction error; risk function.

a: Animalia, P.O. Box 396 - Økern, N-0513 Oslo, Norway

b: Norwegian University of Life Sciences (NMBU), Department of Chemistry, Biotechnology and Food Science, P.O. Box 5003, N-1432 Ås, Norway

1 Introduction and notation

In this paper, we evaluate a linear regression model with a bivariate response variable where one of the responses contains missing data. For practical purposes this situation is likely to occur if predictor variables and one response variable, typically a response variable of subordinate interest, are easily sampled; but the other response, typically the one of primary interest, is hard(er) or more costly to sample. Often, though not necessarily, the fully observed response variable will be a surrogate variable (Upton and Cook, 2014) for the response variable containing missing data. In such situations, a sampling method where the primary response variable is sampled only for a subset of the total sample, might be beneficial, especially if the error terms in the bivariate linear regression model are highly correlated.

When no data is missing and ordinary least squares (OLS) estimators are used, the gain of applying one single multiresponse regression model is limited compared to several single response models, since the regression parameter estimates are equal and unaffected by the covariance structure of the error term. In this paper, we show that the gain of using a bivariate response variable model in cases with missing data for one of the responses might be substantial if the covariance structure of the error terms is taken properly into account.

In the following scalars are denoted by lowercase italic characters, vectors by lowercase bold italic characters and matrices by uppercase bold italic characters. The single elements of any matrix are denoted by the corresponding lowercase italic letter and a subindex, i.e. w_{ij} is the element of the ith row and jth column of $\boldsymbol{W}$. In general, Greek letters are used for parameters and Latin letters are used for random variables. The risk function, $R_{\hat{\boldsymbol{\theta}}}$, for an estimator or predictor, $\hat{\boldsymbol{\theta}}$, with true value $\boldsymbol{\theta}$, is

$$R_{\hat{\boldsymbol{\theta}}} = E\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\right];$$

2 Model spesification

The data are given by an $n_1 \times 2$ matrix of response variables, $\boldsymbol{Y}_1$, and an $n_1 \times p$ matrix of predictor variables, $\boldsymbol{X}_1$, in which the first column is the vector of unity, and the $p-1$ last columns are denoted $\boldsymbol{Z}_1$. If not stated otherwise, $\boldsymbol{Z}_1$ is assumed to be mean centred. The model is a standard bivariate response variable regression model, i.e.

$$\boldsymbol{y}_i^T \sim N_2\left(\boldsymbol{\beta}^T \boldsymbol{x}_i^T,\ \boldsymbol{\Sigma}\right),\ \ i = 1, \ldots, n_1;$$

where $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ denote the ith row of $\boldsymbol{Y}_1$ and $\boldsymbol{X}_1$ respectively. The $p \times 2$ matrix $\boldsymbol{\beta}$, which first and second column are denoted $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, denotes the regression coefficients. The notations $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_z$ are used for the first and $p-1$ last rows of $\boldsymbol{\beta}$ respectively. The 2×2 matrix $\boldsymbol{\Sigma}$ denotes the error covariance matrix, with elements σ_{ij}, $i = 1, 2$, $j = 1, 2$. The model is well known from literature, also in a Bayesian setting (Box and Tiao, 1973; Minka, 2000).

We assume that the data represents a random sample from a larger population, of which a random subsample contains missing data for the second response variable. The observations are rearranged so that the first n_2 ($p < n_2 \leq n_1$) rows of $\boldsymbol{Y}_1$ are fully observed, and for the $n_1 - n_2$ last rows of $\boldsymbol{Y}_1$ only the first column is observed. For the rest of this paper $\boldsymbol{Y}_2$, $\boldsymbol{X}_2$ and $\boldsymbol{Z}_2$ will represent the $n_2 \times 2$, $n_2 \times p$ and $n_2 \times (p-1)$ sub-matrices of the n_2 first rows of $\boldsymbol{Y}_1$, $\boldsymbol{X}_1$ and $\boldsymbol{Z}_1$, respectively. Finally, $\boldsymbol{y}_v$ is the stacked column-vector of the first column of $\boldsymbol{Y}_1$ and the second column of $\boldsymbol{Y}_2$.

The model might be defined in different ways, but the representation

$$\boldsymbol{y}_v \sim N_{n_1+n_2}\left(\boldsymbol{X}_{(+)}\boldsymbol{\beta}_v,\ \ \boldsymbol{\Sigma}_{(+)}\right); \tag{1}$$

where $\boldsymbol{\beta}_v$ is the stacked column-vector of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$, is suitable for the purpose of the rest of this paper. The $(n_1 + n_2) \times (n_1 + n_2)$ covariance matrix $\boldsymbol{\Sigma}_{(+)}$ is the upper left block of $\boldsymbol{\Sigma} \otimes \boldsymbol{I}_{n_1}$. Likewise $\boldsymbol{X}_{(+)}$ denotes the $(n_1 + n_2) \times 2p$ matrix representing the $n_1 + n_2$ first rows of $\boldsymbol{I}_2 \otimes \boldsymbol{X}_1$.

3 Estimators

3.1 The generalized least squares estimator

The $2p \times 1$ vector $\hat{\boldsymbol{\beta}}_v$ denotes the generalized least squares (GLS) estimator based on (1). We denote the first p elements of $\hat{\boldsymbol{\beta}}_v$ by $\hat{\boldsymbol{\beta}}_{v1}$, and the last p elements by $\hat{\boldsymbol{\beta}}_{v2}$. The expression and distribution of $\hat{\boldsymbol{\beta}}_v$ are:

$$\hat{\boldsymbol{\beta}}_v = \left(\boldsymbol{X}_{(+)}^T\boldsymbol{\Sigma}_{(+)}^{-1}\boldsymbol{X}_{(+)}\right)^{-1}\boldsymbol{X}_{(+)}^T\boldsymbol{\Sigma}_{(+)}^{-1}\boldsymbol{y}_v, \quad \hat{\boldsymbol{\beta}}_v \sim N_{2p}\left[\boldsymbol{\beta}_v,\ \ \left(\boldsymbol{X}_{(+)}^T\boldsymbol{\Sigma}_{(+)}^{-1}\boldsymbol{X}_{(+)}\right)^{-1}\right];$$

Remark 1

$\hat{\boldsymbol{\beta}}_{v1}$, equals $\hat{\boldsymbol{\beta}}_{11}$, i.e. the ordinary OLS estimator based on $\boldsymbol{X}_1$.

Remark 2

The expression and distribution for the GLS–estimator for $\boldsymbol{\beta}_2$ is:

$$\hat{\boldsymbol{\beta}}_{v2} = \hat{\boldsymbol{\beta}}_{22} + \sigma_{12}/\sigma_{11}\left(\hat{\boldsymbol{\beta}}_{11} - \hat{\boldsymbol{\beta}}_{21}\right),$$
$$\hat{\boldsymbol{\beta}}_{v2} \sim N_p\left\{\boldsymbol{\beta}_2,\ \sigma_{22}\left(\boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1} - {\sigma_{12}}^2/\sigma_{11}\left[\left(\boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1} - \left(\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\right]\right\};$$

, where $\hat{\boldsymbol{\beta}}_{21}$ and $\hat{\boldsymbol{\beta}}_{22}$ are the standard OLS estimates based on the n_2 first (full) observations for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ respectively. We observe that as n_1 increases towards infinity the distribution for $\hat{\boldsymbol{\beta}}_{v2}$ approaches the distribution of $\hat{\boldsymbol{\beta}}_{22}$ conditional on known $\boldsymbol{\beta}_1$, i.e.

$$\hat{\boldsymbol{\beta}}_{22} \mid \boldsymbol{\beta}_1 = \hat{\boldsymbol{\beta}}_{22} + \sigma_{12}/\sigma_{11}\left(\boldsymbol{\beta}_1 - \hat{\boldsymbol{\beta}}_{21}\right), \quad \hat{\boldsymbol{\beta}}_{22} \mid \boldsymbol{\beta}_1 \sim N_p\left[\boldsymbol{\beta}_2,\ \left(\sigma_{22} - \sigma_{12}^2/\sigma_{11}\right)\left(\boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1}\right];$$

Remark 3

The covariance between $\hat{\boldsymbol{\beta}}_{v1}$ and $\hat{\boldsymbol{\beta}}_{v2}$ is $\sigma_{12}\left(\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}$.

3.2 James-Stein estimator

An alternative estimator for $\boldsymbol{\beta}_v$, $k\hat{\boldsymbol{\beta}}_v$, where $0 \leq k \leq 1$ is known as the James-Stein estimator (Efron and Morris, 1973; James and Stein, 1961). It is well known that if the regularization parameter, k, is set appropriately, then the James–Stein estimator outperforms the GLS–estimator in the sense of having smaller risk, i.e. $R_{k\hat{\boldsymbol{\beta}}_v} \leq R_{\hat{\boldsymbol{\beta}}_v}$.

An obvious objection to the James–Stein estimator is that it is biased for all $k \neq 1$. Another issue is how to set the regularization parameter at a suitable value. Bock (1975) showed how to set k to minimize $R_{k\hat{\boldsymbol{\beta}}_v}$, based on the largest eigenvalue of covariance for $\hat{\boldsymbol{\beta}}_v$. A problem with the estimator $k\hat{\boldsymbol{\beta}}_v$ is that the same regularization, k, is applied to both $\hat{\boldsymbol{\beta}}_{v1}$ and $\hat{\boldsymbol{\beta}}_{v2}$. Brown and Zidek (1980) and Matsuda and Komaki (2015), addressed this problem by analysing different regularization matrices in detail, also based on known error covariance structure. However, they do not deal with missing data for the response variables.

In this paper, we evaluate a variant of the James–Stein estimator, $\tilde{\boldsymbol{\beta}}$, a stacked vector of $\tilde{\boldsymbol{\beta}}_1 = k_1\hat{\boldsymbol{\beta}}_{v1}$ and $\tilde{\boldsymbol{\beta}}_2 = k_2\hat{\boldsymbol{\beta}}_{v2}$ where $0 \leq k_i \leq 1$ for $i = 1, 2$.

4 Prediction error

4.1 General form: No assumptions on predictor variables

Let $\boldsymbol{y}_N$ denote a new observation, and let $\tilde{\boldsymbol{y}}_N$ denote the corresponding predicted value based on the new predictor variable, $\boldsymbol{x}_N$, i.e. a vector of length p where the first element is 1 and the last $p-1$ elements are denoted $\boldsymbol{z}_N$.

Theorem 1: Expected prediction error
The expected squared prediction errors denoted $R_{\tilde{y}Ni}$, for $i = 1, 2$, using the estimator $\tilde{\boldsymbol{\beta}}_i$, and the notation $\boldsymbol{Z}_{2c}$ for the first n_2 rows of $\boldsymbol{Z}_1$ centred with respect to the column means of the same rows, are:

$$\begin{aligned} R_{\tilde{y}_{N1}} =& \sigma_{11} + \sigma_{11}k_1^2\left[1/n_1 + \boldsymbol{z}_N^T\left(\boldsymbol{Z}_1^T\boldsymbol{Z}_1\right)^{-1}\boldsymbol{z}_N\right] + (1-k_1)^2\boldsymbol{x}_N^T\boldsymbol{\beta}_1\boldsymbol{\beta}_1^T\boldsymbol{x}_N; \\ R_{\tilde{y}_{N2}} =& \sigma_{22} + \left(\sigma_{22} - \sigma_{12}^2/\sigma_{11}\right)k_2^2\left[1/n_2 + \boldsymbol{z}_N^T\left(\boldsymbol{Z}_{2c}^T\boldsymbol{Z}_{2c}\right)^{-1}\boldsymbol{z}_N\right] + \\ & \sigma_{12}^2/\sigma_{11}k_2^2\left[1/n_1 + \boldsymbol{z}_N^T\left(\boldsymbol{Z}_1^T\boldsymbol{Z}_1\right)^{-1}\boldsymbol{z}_N\right] + (1-k_2)^2\boldsymbol{x}_N^T\boldsymbol{\beta}_2\boldsymbol{\beta}_2^T\boldsymbol{x}_N; \end{aligned} \tag{2}$$

The proof is deferred to the Appendix.

4.2 Multivariate normal distributed predictor variables

The formulas given by (2) is valid for a new and observed observation $\boldsymbol{x}_N$ and the given calibration set $\boldsymbol{X}_1$. A more general statement about prediction error is the expected squared prediction error over all calibration samples and new observations, $E_{\boldsymbol{x}}\left(R_{\tilde{y}_{Ni}}\right)$, which can be obtained under certain assumptions (Helland and Almøy, 1994).

Assume that all rows of the original (non-centred) $\boldsymbol{Z}_1$, and the new (non-centred) observation, $\boldsymbol{z}_N$, are independent multivariate normal distributed with fixed expectation parameter and a fixed covariance matrix, $\boldsymbol{\Gamma}$. Under these assumptions the expected prediction risks might be given as functions of $\boldsymbol{\Sigma}$, n_i, k_i, β_{0i} and R_i^2, where R_i^2 is the population coefficients of determination.

The natural choices for k_i, denoted $k_{Oracle\ i}$, for $i = 1, 2$, are the values that minimize the expected squared prediction error. The sub-indexing "Oracle" is used in line with Wasserman (2006) and reflects that these values are unattainable in most practical situations.

To compare the precision of different estimators we use the expected ratios of the expected squared prediction errors of the estimators. Since the ratio of expectation is constant,

this ratio is equal to $E_{\boldsymbol{x}}\left(R_{\tilde{y}_{Ni}}\right)/E_{\boldsymbol{x}}\left(R_{\hat{y}_{Ni}}\right)$, where the subscripts $\tilde{y}_{Ni}$ and $\hat{y}_{Ni}$ indicate the estimator.

The expressions given in (3) simplify equations and increase readability. However, they might also be given some kind of interpretation as n_{ei} increases with sample size and decreases as p increases. Further n_{e2} increases as ρ^2 increases, an effect that might be given the interpretation as increased population size for estimating $\boldsymbol{\beta}_2$ by borrowing strength from the observations with missing data.

The constants, c_{i2}, $i = 1, 2$, are functions of the population coefficient of determination, R_i^2, and the intercept term, β_{0i}. The relationship between these constants and both input arguments are positive, though not linear.

$$\begin{aligned} c_{i1} =& (n_i p - 2)/[n_i\left(n_i - p - 1\right)], \quad c_{i2} = \left(n_1 + 1\right) R_i^2/\left[n_1\left(1 - R_i^2\right)\right] + \beta_{0i}^2, \;\; i = 1, 2 \\ n_{e1} =& 1/c_{11}, \quad n_{e2} = \sigma_{22}/\left[\left(\sigma_{22} - \sigma_{12}^2/\sigma_{11}\right) c_{21} + \left(\sigma_{12}^2/\sigma_{11}\right) c_{11}\right]; \end{aligned} \tag{3}$$

Theorem 2: Prediction error over all calibration samples
The expected squared prediction errors over all calibration samples and new observations using the estimator $\tilde{\boldsymbol{\beta}}_i$, for $i = 1, 2$ under the assumptions specified above, are:

$$E_{\boldsymbol{x}}\left(R_{\tilde{y}_{Ni}}\right) = \sigma_{ii}\left[1 + k_i^2 n_{ei}^{-1} + \left(1 - k_i\right)^2 c_{i2}\right]; \tag{4}$$

The proof is deferred to the Appendix.

Corollary 1:
The values for k_i minimizing the expected squared prediction error, and the corresponding expected risk functions are:

$$k_{Oracle\ i} = c_{i2}/\left(n_{ei}^{-1} + c_{i2}\right), \quad E_{\boldsymbol{x}}\left(R_{\tilde{y}_{Ni}}\right)_{Oracle} = \sigma_{ii}\left\{1 + c_{i2}/\left[n_{ei}\left(n_{ei}^{-1} + c_{i2}\right)\right]\right\};$$

Corollary 2:
The values $k_{lim\ i} = \left(n_{ei}^{-1} - c_{i2}\right)/\left(n_{ei}^{-1} + c_{i2}\right)$, has the property that for $k_{lim\ i} < k_i < 1$, then $E_{\boldsymbol{x}}\left(R_{\tilde{y}_{Ni}}\right) < E_{\boldsymbol{x}}\left(R_{\hat{y}_{Ni}}\right)$, where $\hat{y}_{Ni}$ denotes the prediction based on $\hat{\boldsymbol{\beta}}_{vi}$.

Corollary 3:
The expected ratios of the expected squared prediction errors of the estimator $\tilde{\boldsymbol{\beta}}_i$ and the

two competitors $\hat{\boldsymbol{\beta}}_{vi}$ *and* $\hat{\boldsymbol{\beta}}_{2i}$*, note that for* $i = 1$ *those are equal, the ratios of the expected prediction risks, are:*

$$\begin{aligned} E_{\boldsymbol{x}}\left(R_{\tilde{y}_{Ni}}\right)/E_{\boldsymbol{x}}\left(R_{\hat{y}_{Ni}}\right) =& 1 - n_{ei}^{-2}/\left[\left(n_{ei}^{-1}+1\right)\left(n_{ei}^{-1}+c_{i2}\right)\right], \\ E_{\boldsymbol{x}}\left(R_{\tilde{y}_{Ni}}\right)/E_{\boldsymbol{x}}\left(R_{\hat{y}_{2Ni}}\right) =& 1 - \left(c_{i1}n_{ei}^{-1}+c_{i2}n_{ei}^{-1}-c_{i1}c_{i2}\right)/\left[\left(c_{i1}+1\right)\left(n_{ei}^{-1}+c_{i2}\right)\right]; \end{aligned}$$

Corollary 4:

The expected ratio of the expected squared prediction errors based on the estimators $\hat{\boldsymbol{\beta}}_{vi}$ *and* $\hat{\boldsymbol{\beta}}_{2i}$ *equals* 1 *for* $i = 1$ *and has a specially nice expression for* $i = 2$*, where* $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}}$ *is the correlation between the error terms.*

$$E_{\boldsymbol{x}}\left(R_{\hat{y}_{Ni}}\right)/E_{\boldsymbol{x}}\left(R_{\hat{y}_{2Ni}}\right) = 1 - \rho^2(c_{21}-c_{11})/(1+c_{21});$$

Even though, in general, we assume $\boldsymbol{\Sigma}$ to be known in this paper the prediction risk for the estimator

$$\hat{\boldsymbol{\beta}}_{Q2} = \hat{\boldsymbol{\beta}}_{22} + q_{12}/q_{11}\left(\hat{\boldsymbol{\beta}}_{11} - \hat{\boldsymbol{\beta}}_{21}\right), \quad \boldsymbol{Q} = \left[\boldsymbol{Y}_2 - \boldsymbol{X}_2\left(\hat{\boldsymbol{\beta}}_{21}\ \ \hat{\boldsymbol{\beta}}_{22}\right)\right]^T\left[\boldsymbol{Y}_2 - \boldsymbol{X}_2\left(\hat{\boldsymbol{\beta}}_{21}\ \ \hat{\boldsymbol{\beta}}_{22}\right)\right];$$

might be analysed analytically. The fraction q_{12}/q_{11} is an unbiased estimator for σ_{12}/σ_{11} as can be derived using Giri (2003).

Lemma 1:

The expected risk function for prediction error, i.e. $E_{\boldsymbol{x}}\left(R_{\hat{y}_{QN2}}\right)$*, using* $\hat{\boldsymbol{\beta}}_{Q2}$ *as estimator for* $\boldsymbol{\beta}_2$ *is:*

$$E_{\boldsymbol{x}}\left(R_{\hat{y}_{QN2}}\right) = \sigma_{22}\left\{(1+c_{21}) - (c_{21}-c_{11})\left[\rho^2 - \left(1-\rho^2\right)/(n_2-p-2)\right]\right\}; \tag{5}$$

The proof is deferred to the Appendix.

Corollary 5:

The expected ratio of the expected squared prediction errors based on the estimators $\hat{\boldsymbol{\beta}}_{Q2}$ *and* $\hat{\boldsymbol{\beta}}_{22}$ *is:*

$$E_{\boldsymbol{x}}\left(R_{\hat{y}_{QN2}}\right)/E_{\boldsymbol{x}}\left(R_{\hat{y}_{2Ni}}\right) = 1 - \left[(c_{21}-c_{11})(1+c_{21})\right]\left[\rho^2 - \left(1-\rho^2\right)/(n_2-p-2)\right];$$

Do note that if $\rho^2 < 1/(n_2-p-1)$, then the expected prediction risk using the standard OLS estimator, $\hat{\boldsymbol{\beta}}_{22}$, is smaller than using the estimator $\hat{\boldsymbol{\beta}}_{Q2}$.

5 Results

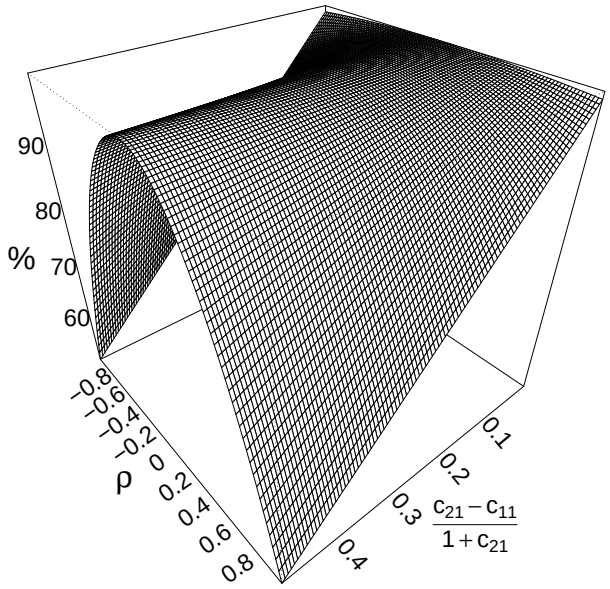


Figure 1: The expected relative size (in %) for the expected squared prediction errors using the estimator $\hat{\boldsymbol{\beta}}_{v2}$ compared with $\hat{\boldsymbol{\beta}}_{22}$ as a function of ρ and the fraction $c_{21} - c_{11}/1 + c_{21}$. As this fraction decreases, it basically means that the relative difference between the sample-sizes n_1 and n_2 also decreases.

As shown in Figure 1, the gain of using $\hat{\boldsymbol{\beta}}_{v2}$ over $\hat{\boldsymbol{\beta}}_{22}$ may be substantial, especially for combinations when ρ^2 and $(c_{21} - c_{11})/(1 + c_{21})$, are both high. The latter expression is basically reflecting the relative difference between n_1 and n_2. Figure 2 shows that further improvements might be achieved by substituting $\hat{\boldsymbol{\beta}}_{vi}$ with $\tilde{\boldsymbol{\beta}}_i$, for $i = 1, 2$ in situations when c_{i2}, basically reflecting the size of R_i^2, is small. The effect diminishes when n_{ei} increases, i.e. when the effective sample size is large.

The results of (2) and (5) were validated via simulations using the software "R" (R Core Team, 2014) and an extension of the package "simrel" (Sæbø, 2015; Sæbø et al., 2015), capable of producing a bivariate response variable. Figure 3 shows different simulation

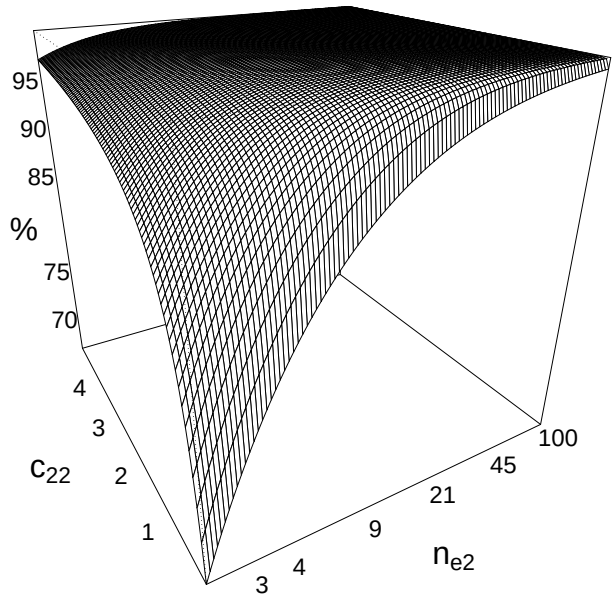


Figure 2: The expected relative size (in %) for the expected squared prediction errors using the estimator $\tilde{\boldsymbol{\beta}}_2$ compared with $\hat{\boldsymbol{\beta}}_{v2}$ as a function of n_{e2} and c_{22}. c_{22} is basically a function of R_2^2 and increases when R_2^2 increases.

tests. As this study is not a simulation study, we contented ourselves with simulating results for a bundle of combinations for n_1 and n_2, varied ρ and plotted the results onto the theoretical risks like shown in Figure 3 for visual validation.

6 Discussion

Our major finding in this study is to show that for linear regression with bivariate response including missing data, there exists an unbiased GLS–estimator, $\hat{\boldsymbol{\beta}}_{v2}$, which reduces the expected prediction error compared with the standard OLS estimator, when the covariance structure of error terms is assumed to be known. The prediction precision might be further improved by shrinking the GLS–estimator by the principles outlined by James and Stein

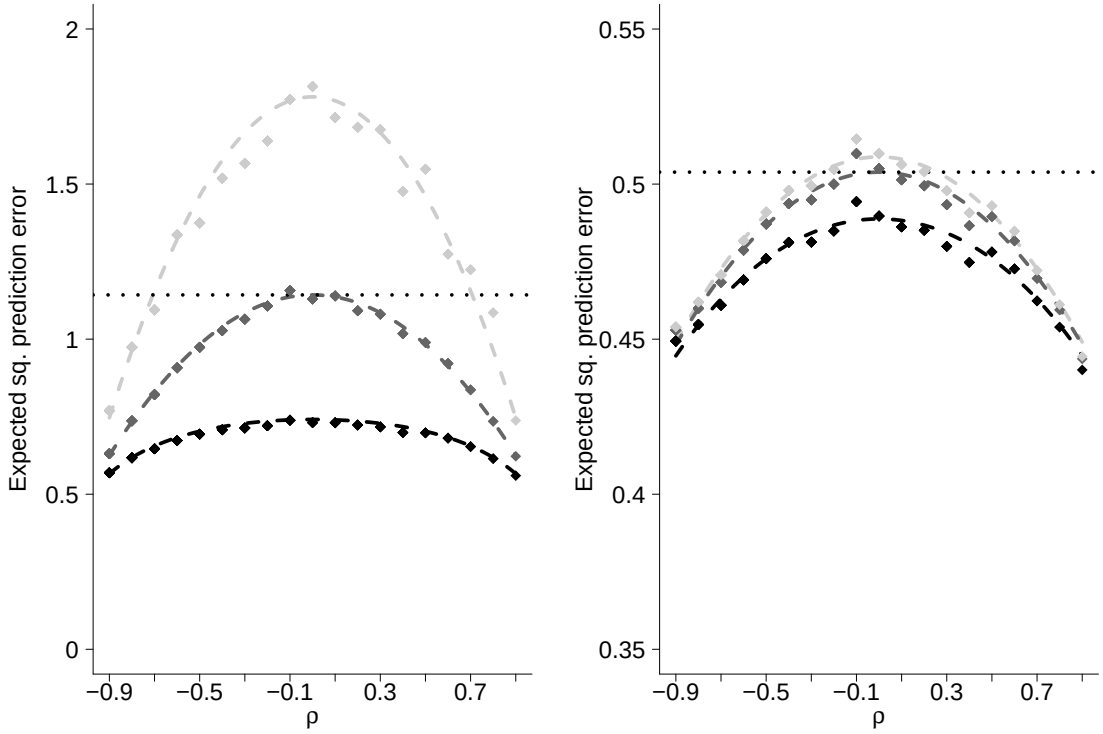


Figure 3: Expected squared prediction error using the predictors $\tilde{\boldsymbol{\beta}}_2$ (black dashed lines) $\hat{\boldsymbol{\beta}}_{v2}$ (dark gray dashed lines), $\hat{\boldsymbol{\beta}}_{Q2}$ (light gray dashed lines) and $\hat{\boldsymbol{\beta}}_{22}$ (black dotted line) as functions of ρ. n_1 equal to 20 in left panel and 50 in right panel, and n_2 equal to 7 in left panel and 20 in right panel. The simulated means are shown by diamonds in the colors corresponding to the theoretical lines. The simulation means are based on 5×10^3 independent calibration sets for each of $\rho = -0.9, -0.8, \ldots, 0.9$, and the estimates from each calibration set is used to predict 10^3 new independent observations. For all panels and simulations $p = 4$, $R_1^2 = 0.4$, $R_2^2 = 0.6$ and $\beta_{01} = \beta_{02} = 0$.

(1961).

The natural next step is to test out estimators that do not assume known covariance structure ($\boldsymbol{\Sigma}$) and known coefficients of determination (R_i^2). The obvious choice corresponding to $\tilde{\boldsymbol{\beta}}$ is some kind of empirical Bayes estimator. Their connection to the James–Stein estimator is well documented by a series of papers by Efron and Morris (1971, 1972a,b, 1973, 1975, 1976). Other candidates would be restricted maximum likelihood estimators, corresponding to $\hat{\boldsymbol{\beta}}_v$, and possibly an extension of the (C)PLS estimator (Indahl et al., 2009), capable of utilizing information from observations with missing data.

The generalisation of assuming predictors to be multivariate normal distributed might be severely biased in a lot of practical situations, especially when experiments are designed. However, for many, perhaps the majority, of practical situations, the assumption might be justified at least after some normalizing transformation of variables. The validity of

the theoretical results when the assumption of normally distributed predictors is violated, has been tested for the OLS estimators by simulating results using randomly distributed, not normal distributed predictors. The effect of non-normality was found to be negligible. This seems intuitively correct, as the principles of the central limit theorem should also be applicable for the current situation.

References

Bock, M. (1975). Minimax Estimators of the Mean of a Multivariate Normal Distribution. *Ann. Statist.*, 3:209–218.

Box, G. and Tiao, G. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley series in behavioral science: quantitative methods. Addison-Wesley Pub. Co.

Brown, P. J. and Zidek, J. V. (1980). Adaptive Multivariate Ridge Regression. *Ann. Statist.*, 8:64–74.

Efron, B. and Morris, C. (1971). Limiting the Risk of Bayes and Empirical Bayes Estimators-Part I: The Bayes Case. *J. Amer. Statist. Assoc.*, 66(336):807–815.

Efron, B. and Morris, C. (1972a). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika*, 59(2):335–347.

Efron, B. and Morris, C. (1972b). Limiting the Risk of Bayes and Empirical Bayes Estimators-Part II: The Empirical Bayes Case. *J. Amer. Statist. Assoc.*, 67(337):130–139.

Efron, B. and Morris, C. (1973). Stein's Estimation Rule and its Competitors-An Empirical Bayes Approach. *J. Amer. Statist. Assoc.*, 68(341):117–130.

Efron, B. and Morris, C. (1975). Data Analysis Using Stein's Estimator and its Generalizations. *J. Amer. Statist. Assoc.*, 70(350):311–319.

Efron, B. and Morris, C. (1976). Families of Minimax Estimators of the Mean of a Multivariate Normal Distribution. *Ann. Statist.*, 4:11–21.

Giri, N. C. (2003). *Multivariate Statistical Analysis: Revised and Expanded*, volume 171. CRC Press.

Helland, I. S. and Almøy, T. (1994). Comparison of Prediction Methods when Only a Few Components are Relevant. *J. Amer. Statist. Assoc.*, 89(426):583–591.

Indahl, U. G., Liland, K. H., and Næs, T. (2009). Canonical partial least squares - a unified PLS approach to classification and regression problems. *Journal of Chemometrics*, 23(9):495–504.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). *Multivariate analysis*. Academic press.

Matsuda, T. and Komaki, F. (2015). Singular value shrinkage priors for Bayesian prediction. *Biometrica*, 102(4):843–854.

Minka, T. (2000). Bayesian linear regression. Technical report, Microsoft Research Cambridge.

Petersen, K. B. and Pedersen, M. S. (2012). The matrix cookbook. Technical report, University of Waterloo.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sæbø, S. (2015). *simrel: Linear Model Data Simulation and Design of Computer Experiments*. R package version 1.1-0.

Sæbø, S., Almøy, T., and Helland, I. S. (2015). simrel-A versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors. *Chemometrics and Intelligent Laboratory Systems*.

Upton, G. and Cook, I. (2014). *A Dictionary of Statistics 3e*. Oxford university press.

Wasserman, L. (2006). *All of Nonparametric Statistics*. Springer Science & Business Media.

A Appendix

A.1 Proof of Theorem 1

Since $\boldsymbol{x}_N$ is centred we may write $R_{\tilde{y}_{Ni}} = \sigma_{ii} + R_{\boldsymbol{x}_N^T\tilde{\boldsymbol{\beta}}_i}$. Further since $\hat{\boldsymbol{\beta}}_{vi}$ is normally distributed, so are $\tilde{\boldsymbol{\beta}}_i$ and $\boldsymbol{x}_n^T\left(\boldsymbol{\beta}_i - \tilde{\boldsymbol{\beta}}_i\right)$ for $i = 1, 2$. We have

$$E\left[\boldsymbol{x}_N^T\left(\boldsymbol{\beta}_i - \tilde{\boldsymbol{\beta}}_i\right)\right] = (1 - k_i)\,\boldsymbol{x}_n^T\boldsymbol{\beta}_i, \quad var\left[\boldsymbol{x}_n^T\left(\boldsymbol{\beta}_i - \tilde{\boldsymbol{\beta}}_i\right)\right] = k_i^2\boldsymbol{x}_N^T\left(\boldsymbol{X}_{(+)}^T\boldsymbol{\Sigma}_{(+)}^{-1}\boldsymbol{X}_{(+)}\right)^{-1}\boldsymbol{x}_N;$$

It might be shown that:

$$\boldsymbol{x}_N^T \left(\boldsymbol{X}_1^T \boldsymbol{X}_1\right)^{-1} \boldsymbol{x}_N = 1/n_1 + \boldsymbol{z}_N^T \left(\boldsymbol{Z}_1^T \boldsymbol{Z}_1\right)^{-1} \boldsymbol{z}_N, \quad \boldsymbol{x}_N^T \left(\boldsymbol{X}_2^T \boldsymbol{X}_2\right)^{-1} \boldsymbol{x}_N = 1/n_2 + \boldsymbol{z}_N^T \left(\boldsymbol{Z}_{2c}^T \boldsymbol{Z}_{2c}\right)^{-1} \boldsymbol{z}_N;$$

, where both $\boldsymbol{Z}_{2c}$ and $\boldsymbol{z}_N$ are centred with respect to the n_2 first rows, and $\boldsymbol{X}_2$ is centred with respect to all n_1 rows. Then, since

$$R_{\boldsymbol{x}_N^T \tilde{\boldsymbol{\beta}}_i} = var\left[\boldsymbol{x}_N^T \left(\boldsymbol{\beta}_i - \tilde{\boldsymbol{\beta}}_i\right)\right] + \left\{E\left[\boldsymbol{x}_N^T \left(\boldsymbol{\beta}_i - \tilde{\boldsymbol{\beta}}_i\right)\right]\right\}^2;$$

we get (2).

A.2 Proof of Theorem 2

By applying the rules for double expectations, we may write $E_{\boldsymbol{x}}\left(R_{\tilde{y}_{Ni}}\right) = E_{\boldsymbol{X}_1}\left[E_{\boldsymbol{x}_N|\boldsymbol{X}_1}\left(R_{\tilde{y}_{Ni}}\right)\right]$. Due to the assumption of independent normal distribution of the rows of $\boldsymbol{Z}_1$, we have that $\boldsymbol{Z}_1^T \boldsymbol{Z}_1$ and $\boldsymbol{Z}_{2c}^T \boldsymbol{Z}_{2c}$ are two Wishard distributed variables with scale matrix $\boldsymbol{\Gamma}^{-1}$ and $n_1 - 1$ and $n_2 - 1$ degrees of freedom, respectively (Mardia et al., 1979). Thus, their inverse matrices are inverse Wishard distributed with the same parameters. Due to the centring of $\boldsymbol{z}_N$, we have that $\boldsymbol{z}_N$ is multivariate normally distributed with zero mean and covariance matrix $\left[(n_i + 1)/n_i\right]\boldsymbol{\Gamma}$ when centred using all rows ($i = 1$) or just the n_2 first rows ($i = 2$).

By using rules for quadratic terms (Petersen and Pedersen, 2012), and the rules for expectation of the trace, we find

$$E\left[\boldsymbol{z}_N^T \left(\boldsymbol{Z}_1^T \boldsymbol{Z}_1\right)^{-1} \boldsymbol{z}_N\right] = \left[(n_1 + 1)/n_1\right]\left[(p-1)/(n_1 - p - 1)\right] = c_{11} - 1/n_1;$$
$$E\left[\boldsymbol{z}_N^T \left(\boldsymbol{Z}_{2c}^T \boldsymbol{Z}_{2c}\right)^{-1} \boldsymbol{z}_N\right] = \left[(n_2 + 1)/n_2\right]\left[(p-1)/(n_2 - p - 1)\right] = c_{21} - 1/n_2;$$

Further, $E\left(\boldsymbol{x}_N^T \boldsymbol{\beta}_i \boldsymbol{\beta}_i^T \boldsymbol{x}_N\right) = E\left(\beta_{0i}^2\right) + E\left(\boldsymbol{z}_N^T \boldsymbol{\beta}_{zi} \boldsymbol{\beta}_{zi}^T \boldsymbol{z}_N\right)$ since $E\left(\boldsymbol{z}_N\right) = \boldsymbol{0}_{(p-1)}$ for $i = 1, 2$. Finally, the term $E\left(\boldsymbol{z}_N^T \boldsymbol{\beta}_{zi} \boldsymbol{\beta}_{zi}^T \boldsymbol{z}_N\right) = \sigma_{ii} R_i^2/(1 - R_i^2)$ is given by definition.

Then (4) is obtained by substituting the elements in the expressions in (2) by the general terms shown above. The corollaries are given without further proof as they are easily derived mostly by minimizing functions with respect to k_1 and k_2.

A.3 Proof of Lemma 1

Conditional on known $\boldsymbol{Q}$ it might be shown by matrix algebra and the means of the multivariate normal distribution that the distribution of $\hat{\boldsymbol{\beta}}_{Q2}$ is:

$$\hat{\boldsymbol{\beta}}_{Q2} \sim N_p \left\{ \boldsymbol{\beta}_2,\ \sigma_{22} \left(\boldsymbol{X}_2^T \boldsymbol{X}_2 \right)^{-1} - \left(2\sigma_{12} q_{12}/q_{11} - \sigma_{11} q_{12}^2/q_{11}^2 \right) \left[\left(\boldsymbol{X}_2^T \boldsymbol{X}_2 \right)^{-1} - \left(\boldsymbol{X}_1^T \boldsymbol{X}_1 \right)^{-1} \right] \right\};$$

Due to the properties of the normal–gamma distribution, we know that (Giri, 2003):

$$E\left(q_{12}/q_{11}\right) = \sigma_{12}/\sigma_{11}, \quad E\left(q_{12}^2/q_{11}^2\right) = \sigma_{12}^2/\sigma_{11}^2 + |\boldsymbol{\Sigma}|/\left(\sigma_{11}^2 \left(n_2 - p - 2\right)\right);$$

Then, since $var_Q \left[E\left(\hat{\boldsymbol{\beta}}_{Q2} \right) \right] = \mathbf{0}_p^T \mathbf{0}_p$ we find $var\left(\hat{\boldsymbol{\beta}}_{Q2} \right) = E_Q \left[var\left(\hat{\boldsymbol{\beta}}_{Q2} \right) \right]$, leading to (5).

Paper II

Linear regression with bivariate response variable containing missing data. An empirical Bayes strategy to increase prediction precision.

Lars Erik Gangsei[a,b], Trygve Almøy[b] & Solve Sæbø[b]

Abstract: Methods for linear regression with multivariate response variables are well described in statistical literature, including in a Bayesian setting. The state–of–the–art Bayesian methods including conjugate prior distributions require a set of data containing no missing values. In this study we present a conjugate prior distribution, highly motivated by the normal–inverse–Wishart distribution, for bivariate linear regression, where one of the response variables contains missing data. Further we demonstrate how an empirical Bayes approach improves prediction precision compared with standard methods in a dominant share of practical cases, and that the improvement under certain, yet plausible, conditions is substantial.

Keywords: bivariate linear regression; James–Stein estimator; missing data; prediction error; conjugate prior.

a: Animalia, P.O. Box 396 - Økern, N-0513 Oslo, Norway

b: Norwegian University of Life Sciences (NMBU), Department of Chemistry, Biotechnology and Food Science, P.O. Box 5003, N-1432 Ås, Norway

1 Introduction and notation

In this paper, we evaluate different estimators for the regression parameters in linear regression with bivariate response variable. The estimators are based on data in which one of the responses contains missing data. Gangsei et al. (2016) showed that under the assumption of known error covariance structure, the prediction precision for the variable containing missing data might be substantially improved by using the generalized least squares (GLS) estimator rather than the ordinary least squares (OLS) estimator. Further additional improvement might be reached by substituting the GLS–estimator with a James–Stein inspired estimator (James and Stein, 1961), i.e. a biased shrinkage estimator.

The data structure with missing response data is likely to occur in a wide range of practical situations. The response without missing data might be viewed as a surrogate variable (Upton and Cook, 2014), i.e. as a cheaper/ easier alternative measurement for the response, or it might simply be viewed as a predictor variable that will be absent for future predictions.

In this paper, we deal primarily with a Bayesian method for analysing such data. The novel part is to show that there, conditional on a suitable prior distribution, exits a closed form solution for the posterior distribution, even if one of the response variables contains missing data. In turn, this enables us to find the expression for the model evidence (also known as the marginal likelihood) (DeGroot, 2005, chap.9). The model evidence might be used for implementation of Bayesian model selection (Kass and Raftery, 1995).

In addition we include the perspective of the classic results obtained in a series of papers by Efron and Morris (1971, 1972a,b, 1973, 1975, 1976), and show how an empirical Bayes strategy results in a regression parameter estimate that might be viewed as a James-Stein estimator very similar to the estimator evaluated by Gangsei et al. (2016).

The theoretical results are tested via a simulation study. The simulations are used to evaluate aspects of the model for which we lack analytical results. Both the analytical analysis and the simulation study are focused on evaluation of prediction precision.

In the following, scalars are denoted by lowercase italic characters, vectors by lowercase bold italic characters and matrices by uppercase bold italic characters. The single elements of any matrix are denoted by the corresponding lowercase italic letter and a subindex, i.e. w_{ij} is the element of the ith row and jth column of $\boldsymbol{W}$. In general, Greek letters are used for parameters and Latin letters are used for random variables. The risk function, $R_{\hat{\boldsymbol{\theta}}}$, for an estimator or predictor, $\hat{\boldsymbol{\theta}}$, with true value $\boldsymbol{\theta}$, is

$$R_{\hat{\boldsymbol{\theta}}} = E\left[\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)^T \left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}\right)\right];$$

2 Likelihood

The model, that is, the likelihood, is similar to the model evaluated by Gangsei et al. (2016), and we use the same notations. The basic model, without missing data, is well known from previous work, also in a Bayesian setting (Box and Tiao, 1973; Minka, 2000). The basic properties of the likelihood are reiterated in the following paragraphs.

The data are given by a $n_1 \times 2$ matrix of response variables, $\boldsymbol{Y}_1$, and a $n_1 \times p$ matrix of predictor variables, $\boldsymbol{X}_1$. The model is a standard bivariate response regression model, i.e.

$$\boldsymbol{y}_i^T \overset{i.i.d.}{\sim} N_2\left(\boldsymbol{\beta}^T \boldsymbol{x}_i^T,\ \boldsymbol{\Sigma}\right),\ i = 1, \ldots, n_1$$

where $\boldsymbol{y}_i$ and $\boldsymbol{x}_i$ denote the ith row of $\boldsymbol{Y}_1$ and $\boldsymbol{X}_1$, respectively. The $p \times 2$ matrix $\boldsymbol{\beta}$ holds the regression coefficients and the 2×2 matrix $\boldsymbol{\Sigma}$ is the error covariance matrix.

We assume that the data represents a random sample from a larger population, of which a random subsample contains missing data for the second response variable. The observations are rearranged so that the first n_2 ($p < n_2 \leq n_1$) rows of $\boldsymbol{Y}_1$ are fully observed, and for the $n_1 - n_2$ last rows of $\boldsymbol{Y}_1$ only the first column is observed. For the rest of this paper $\boldsymbol{Y}_2$ and $\boldsymbol{X}_2$ will represent the fully observed $n_2 \times 2$ and $n_2 \times p$ sub-matrices of $\boldsymbol{Y}_1$ and $\boldsymbol{X}_1$ respectively. Further $\boldsymbol{y}_{11}$ denotes the first column of $\boldsymbol{Y}_1$, $\boldsymbol{y}_{21}$ and $\boldsymbol{y}_{22}$ denotes the first and second column of $\boldsymbol{Y}_2$.

3 Full Bayesian analysis - conjugate prior

In order to get an easier expression for the prior and posterior distributions we introduce a parameter transformation $\boldsymbol{\Sigma}\left(\sigma_{11}, \sigma_{22}, \sigma_{12}\right) \rightarrow \boldsymbol{\Lambda}\left(\lambda_{11}, \lambda_{22}, \lambda_{12}\right)$, inspired by Giri (2003, chap. 6). The new parameters are given by $\lambda_{11} = 1/\sigma_{11}$, $\lambda_{22} = \sigma_{11}/|\boldsymbol{\Sigma}|$ and $\lambda_{12} = -\sigma_{12}/\sigma_{11}$. Gangsei et al. (2016) showed that the GLS–estimate for regression parameters is:

$$\hat{\boldsymbol{\beta}}_v = \left(\hat{\boldsymbol{\beta}}_{v1}^T\ \ \hat{\boldsymbol{\beta}}_{v2}^T\right)^T = \left\{\hat{\boldsymbol{\beta}}_{11}^T\ \ \left[\hat{\boldsymbol{\beta}}_{22} + \lambda_{12}\left(\hat{\boldsymbol{\beta}}_{21} - \hat{\boldsymbol{\beta}}_{11}\right)\right]^T\right\}^T;$$

, where $\hat{\boldsymbol{\beta}}_{11}$ denotes the standard OLS–estimator for $\boldsymbol{\beta}_1$ based on $\boldsymbol{X}_1$, and $\hat{\boldsymbol{\beta}}_{21}$ and $\hat{\boldsymbol{\beta}}_{22}$ are the standard OLS estimates for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ respectively, based on $\boldsymbol{X}_2$, i.e. the n_2 first (complete) observations.

The hyperparameters used in the prior distribution are given by the symmetric positive definite 2×2 matrix $\boldsymbol{\Phi}$, the scalar ϕ_3, the two symmetric positive definite $p \times p$ matrices $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Psi}_2$, the 2×1 vector $\boldsymbol{\zeta}$ and the $p \times 2$ matrix $\boldsymbol{\beta}_0$. $\boldsymbol{\beta}_{0v}$ denotes the stacked column vector of $\boldsymbol{\beta}_0$. Finally let $\boldsymbol{\Omega}_0$, $\boldsymbol{\Omega}_X$,$\boldsymbol{\Upsilon}$, $\tilde{\boldsymbol{\Omega}}$, υ_3 and $\tilde{\boldsymbol{\beta}}_v$ be given by:

$$
\begin{aligned}
\boldsymbol{\Omega}_0 &= \lambda_{11}^{-1}\left[(1 \quad -\lambda_{12})^T(1 \quad -\lambda_{12})\right] \otimes \boldsymbol{\Psi}_1^{-1} + \lambda_{22}^{-1}\left[(0 \quad 1)^T(0 \quad 1)\right] \otimes \boldsymbol{\Psi}_2^{-1};\\
\boldsymbol{\Omega}_x &= \lambda_{11}^{-1}\left[(1 \quad -\lambda_{12})^T(1 \quad -\lambda_{12})\right] \otimes \left(\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1} + \lambda_{22}^{-1}\left[(0 \quad 1)^T(0 \quad 1)\right] \otimes \left(\boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1};\\
\tilde{\boldsymbol{\Omega}} &= \lambda_{11}^{-1}\left[(1 \quad -\lambda_{12})^T(1 \quad -\lambda_{12})\right] \otimes \left(\boldsymbol{\Psi}_1 + \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1} + \lambda_{22}^{-1}\left[(0 \quad 1)^T(0 \quad 1)\right] \otimes \left(\boldsymbol{\Psi}_2 + \boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1};\\
\boldsymbol{\Upsilon} &= \left(\boldsymbol{Y}_2 - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\right)^T\left(\boldsymbol{Y}_2 - \boldsymbol{X}_2\hat{\boldsymbol{\beta}}_2\right) + \left(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_0\right)^T\left[\left(\boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1} + \boldsymbol{\Psi}_2^{-1}\right]\left(\hat{\boldsymbol{\beta}}_2 - \boldsymbol{\beta}_0\right) + \boldsymbol{\Phi};\\
\upsilon_3 &= \left(\boldsymbol{y}_1 - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{11}\right)^T\left(\boldsymbol{y}_1 - \boldsymbol{X}_1\hat{\boldsymbol{\beta}}_{11}\right) + \left(\hat{\boldsymbol{\beta}}_{11} - \boldsymbol{\beta}_{01}\right)^T\left[\left(\boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1} + \boldsymbol{\Psi}_1^{-1}\right]\left(\hat{\boldsymbol{\beta}}_{11} - \boldsymbol{\beta}_{01}\right) + \phi_3;\\
\tilde{\boldsymbol{\beta}}_v &= \tilde{\boldsymbol{\Omega}}\left(\boldsymbol{\Omega}_X^{-1}\hat{\boldsymbol{\beta}}_v + \boldsymbol{\Omega}_0^{-1}\boldsymbol{\beta}_{0v}\right) = \left(\tilde{\boldsymbol{\beta}}_1^T \quad \tilde{\boldsymbol{\beta}}_2^T\right)^T;\\
\tilde{\boldsymbol{\beta}}_1 &= \left(\boldsymbol{\Psi}_\mathbf{1} + \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1^T\boldsymbol{y}_1 + \boldsymbol{\Psi}_1\boldsymbol{\beta}_{01}\right);\\
\tilde{\boldsymbol{\beta}}_2 &= \left(\boldsymbol{\Psi}_\mathbf{2} + \boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2^T\boldsymbol{y}_{22} + \boldsymbol{\Psi}_2\boldsymbol{\beta}_{02}\right) + \lambda_{12}\left[\left(\boldsymbol{\Psi}_\mathbf{2} + \boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2^T\boldsymbol{y}_{21} + \boldsymbol{\Psi}_2\boldsymbol{\beta}_{01}\right) - \tilde{\boldsymbol{\beta}}_1\right];
\end{aligned}
$$

Theorem 1 Conjungate prior distribution

Conjugate prior distributions for λ_{11}, λ_{22}, λ_{12} and $\boldsymbol{\beta}$ are

$$
\begin{aligned}
\lambda_{11} &\sim Ga\left(\zeta_1/2,\ \ \phi_3/2\right);\\
\lambda_{22} &\sim Ga\left[\zeta_2/2,\ \ \left(\phi_{22} - \phi_{12}^2\phi_{11}\right)/2\right];\\
\lambda_{12} \mid \lambda_{22} &\sim N\left[-\phi_{12}/\phi_{11},\ \ 1/(\phi_{11}\lambda_{22})\right];\\
\boldsymbol{\beta}_v \mid \boldsymbol{\Lambda} &\sim N_{2p}\left(\boldsymbol{\beta}_{0v},\ \ \boldsymbol{\Omega}_0\right);
\end{aligned}
\tag{1}
$$

For these priors the conditional posterior distributions are

$$
\begin{aligned}
\lambda_{11} \mid \boldsymbol{Y} &\sim Ga\left[(\zeta_1 + n_1)/2,\ \ \upsilon_3/2\right];\\
\lambda_{22} \mid \boldsymbol{Y} &\sim Ga\left[(\zeta_2 + n_2)/2,\ \ \left(\upsilon_{22} - \upsilon_{12}^2\upsilon_{11}\right)/2\right];\\
\lambda_{12} \mid \boldsymbol{Y}, \lambda_{22} &\sim N\left[-\upsilon_{12}/\upsilon_{11},\ \ 1/(\upsilon_{11}\lambda_{22})\right];\\
\boldsymbol{\beta}_v \mid \boldsymbol{Y}, \boldsymbol{\Lambda} &\sim N_{2p}\left(\tilde{\boldsymbol{\beta}}_v,\ \ \tilde{\boldsymbol{\Omega}}\right);
\end{aligned}
\tag{2}
$$

The proof is deferred to the Appendix.

Remark 1
If $\phi_{11} = \phi_3$ and $\zeta_2 = \zeta_1 + 1$, the prior distribution for $\boldsymbol{\Lambda}$ might be represented as $\boldsymbol{\Sigma}^{-1} \sim W\left(\boldsymbol{\Phi}^{-1},\ \zeta_2\right)$.

Remark 2
If $\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2$ the prior distribution for $\boldsymbol{\beta}_v \mid \boldsymbol{\Lambda}$ might be represented as $\boldsymbol{\beta} \mid \boldsymbol{\Lambda} \sim MN_{p,2}\left(\boldsymbol{\beta}_0,\ \boldsymbol{\Psi}_1^{-1},\ \boldsymbol{\Sigma}\right)$. Consequently if $\phi_{11} = \phi_3$, $\zeta_2 = \zeta_1 + 1$ and $\boldsymbol{\Psi}_1 = \boldsymbol{\Psi}_2$ the prior distribution for $(\boldsymbol{\beta}_v, \boldsymbol{\Lambda})$, expressed as $(\boldsymbol{\beta}_v, \boldsymbol{\Sigma}^{-1})$, is given by $(\boldsymbol{\beta}_v, \boldsymbol{\Sigma}^{-1}) \sim MNW\left(\boldsymbol{\beta}_0,\ \boldsymbol{\Psi}_1^{-1},\ \boldsymbol{\Phi}^{-1},\ \zeta_2\right)$.

Since both the prior distribution and the posterior distribution are compound distributions of similar form, independent Monte Carlo estimates for $\boldsymbol{\Lambda}$ and $\boldsymbol{\beta}$, i.e. estimates free of auto correlation, are easily obtained by an iterative algorithm. First values for λ_{11} and λ_{22} are sampled independently, λ_{12} is sampled based on values for λ_{22} and finally $\boldsymbol{\beta}$ is sampled based on $\boldsymbol{\Lambda}$. Posterior means and variances for $\boldsymbol{\beta}$ and $\boldsymbol{\Lambda}$ are given in Appendix A.1.

The model evidence, denoted $\pi\left(\boldsymbol{Y}\right)$, might be used for Bayesian model selection (Kass and Raftery, 1995). In this paper, we highlight the application where model evidence, in combination with a suitable prior distribution, is used for tuning the prior hyperparameters into model regularization parameters in accordance with the principles outlined by James and Stein (1961) and Efron and Morris (1971, 1972a,b, 1973, 1975).

Lemma 1 *The model evidence, $\pi\left(\boldsymbol{Y}\right)$, when the prior distribution in (1) is applied to the model is given by*

$$\pi\left(\boldsymbol{Y}\right) = \prod_{i=1}^{2} \left\{ \frac{\Gamma\left[(n_i + \zeta_i)/2\right]}{\pi^{n_i/2}\Gamma\left(\zeta_i/2\right)} \cdot \frac{\left|\boldsymbol{\Psi}_i\right|^{1/2}}{\left|\boldsymbol{\Psi}_i + \boldsymbol{X}_i^T\boldsymbol{X}_i\right|^{1/2}} \right\} \cdot \frac{\upsilon_{11}^{(n_2+\zeta_2-1)/2}}{\phi_{11}^{(\zeta_2-1)/2}} \cdot \frac{\boldsymbol{\phi}_3^{\zeta_1/2}}{\upsilon_3^{(n_1+\zeta_1)/2}} \cdot \frac{\left|\boldsymbol{\Phi}\right|^{\zeta_2/2}}{\left|\boldsymbol{\Upsilon}\right|^{(n_2+\zeta_2)/2}} \tag{3}$$

The proof is deferred to the Appendix.

4 Empirical Bayes

Inspired by a desire to evaluate a James-Stein estimator using empirical Bayes methods, we applied a prior distribution including four free hyper-parameters given by the vectors $\boldsymbol{\alpha} = \left(\alpha_1 \ \ \alpha_2\right)^T$ and $\boldsymbol{\gamma} = \left(\gamma_1 \ \ \gamma_2\right)^T$ of length 2, a prior that might also be viewed as a variant of Zellners g-prior (Zellner, 1986) and also influenced by Minka (2000).

For the empirical Bayes, we use centred data for non-categorical columns of $\boldsymbol{X}_1$ and both columns of $\boldsymbol{Y}_1$. Note that even though the data are centered, the first row of $\boldsymbol{\beta}$ is the traditional "intercept", i.e. the first column of $\boldsymbol{X}_1$ is $\mathbf{1}_{n_1}$. Due to the centring the first element of $\hat{\boldsymbol{\beta}}_{11}$ equals zero. However the elements in the first row of $\hat{\boldsymbol{\beta}}_{22}$ is not equal to zero as $\boldsymbol{X}_2$ is centred with respect to all n_1 columns.

Proposition 1 - Empirical Bayes prior distribution
We propose to use the following prior parameters:

$$\boldsymbol{\zeta} = \boldsymbol{\gamma}, \;\; \phi_3 = \gamma_1, \;\; \boldsymbol{\Phi} = \gamma_2 \boldsymbol{I}_2;$$
$$\boldsymbol{\Psi}_i = (\alpha_i/n_i)\left(\boldsymbol{X}_i^T \boldsymbol{X}_i\right), \;\; \boldsymbol{\beta}_0 = \mathbf{0}_p \mathbf{0}_2^T;$$

A nonlinear optimizer was used to set the hyperparameters at values maximizing the logarithm of model evidence, see Appendix A.2.

Remark 3
The posterior means for $\boldsymbol{\beta}$, denoted $\tilde{\boldsymbol{\beta}}_{EB} = \left(\tilde{\boldsymbol{\beta}}_{1\,EB} \;\; \tilde{\boldsymbol{\beta}}_{2\,EB}\right)$ using the empirical Bayes method are:

$$\tilde{\boldsymbol{\beta}}_{1\,EB} = [n_1/(n_1+\alpha_1)]\,\hat{\boldsymbol{\beta}}_{11};$$
$$\tilde{\boldsymbol{\beta}}_{2\,EB} = [n_2/(n_2+\alpha_2)]\,\hat{\boldsymbol{\beta}}_{22} - (\upsilon_{12}/\upsilon_{11})\left\{[n_2/(n_2+\alpha_2)]\,\hat{\boldsymbol{\beta}}_{21} - \tilde{\boldsymbol{\beta}}_{1\,EB}\right\};$$

4.1 Justification for empirical Bayes

We aim at finding an estimator for $\boldsymbol{\beta}_2$ that exploits information from the observations with missing data. As shown by Gangsei et al. (2016), the estimator $\hat{\boldsymbol{\beta}}_{v2}$ has this property in the sense that the expected prediction error using $\hat{\boldsymbol{\beta}}_{v2}$ is smaller than (or equal to in the special case $\rho = 0$) the expected prediction error using the standard OLS estimator, $\hat{\boldsymbol{\beta}}_{22}$.

However, unknown error covariance structure complicates the problem. We let

$$\boldsymbol{Q} = \left[\boldsymbol{Y}_2 - \boldsymbol{X}_2\left(\hat{\boldsymbol{\beta}}_{21} \;\; \hat{\boldsymbol{\beta}}_{22}\right)\right]^T \left[\boldsymbol{Y}_2 - \boldsymbol{X}_2\left(\hat{\boldsymbol{\beta}}_{21} \;\; \hat{\boldsymbol{\beta}}_{22}\right)\right];$$

Gangsei et al. (2016) showed that just replacing λ_{12} in the equation for $\hat{\boldsymbol{\beta}}_{v2}$ by its unbiased estimate q_{12}/q_{11} (Giri, 2003), was beneficial when ρ^2 was large. However, when ρ^2 is small, particularly in combination with small values for $n_2 - p$, this estimator, denoted $\hat{\boldsymbol{\beta}}_{2\hat{\lambda}}$ (in Gangsei et al. (2016) the notation $\hat{\boldsymbol{\beta}}_{Q2}$ was used), performs worse than the ordinary OLS estimator.

A natural candidate as unbiased estimator for $\boldsymbol{\beta}_2$ is $\hat{\boldsymbol{\beta}}_{2\tilde{\lambda}}$, where λ_{12} in the equation for $\hat{\boldsymbol{\beta}}_{v2}$ is replaced by $\tilde{\lambda}_{12} = -v_{12}/v_{11}$. $\tilde{\lambda}_{12}$ is the posterior mean for λ_{12}, and thereby a natural point-estimate for λ_{12}. Note that $(1/\lambda_{22} + \lambda_{12}^2/\lambda_{11}) = \sigma_{22}$. The distribution of $\hat{\boldsymbol{\beta}}_{2\tilde{\lambda}}$ is:

$$\hat{\boldsymbol{\beta}}_{2\tilde{\lambda}} \sim N_p \left\{ \boldsymbol{\beta}_2,\ \left(1/\lambda_{22} + \lambda_{12}^2/\lambda_{11}\right) \left(\boldsymbol{X}_2^T \boldsymbol{X}_2\right)^{-1} - f\left(\boldsymbol{\Lambda}, \tilde{\lambda}_{12}\right) \left[\left(\boldsymbol{X}_2^T \boldsymbol{X}_2\right)^{-1} - \left(\boldsymbol{X}_1^T \boldsymbol{X}_1\right)^{-1} \right] \right\};$$

, where

$$f\left(\boldsymbol{\Lambda}, \tilde{\lambda}_{12}\right) = \left(2\lambda_{12}\tilde{\lambda}_{12}\right)/\lambda_{11} - \tilde{\lambda}_{12}^2/\lambda_{11};$$

Examination of the expression for $\boldsymbol{\Upsilon}$ clearly shows that γ_2, conditional on known α_2, might be set in a way such that $\tilde{\lambda}_{12} = -a \cdot q_{12}/q_{11}$, where $0 \leq a \leq 1$. Thus γ_2 might be seen as a shrinkage parameter, having the effect of shrinking the estimate for λ_{12} towards zero.

By general theory of the model in question, we know that $\boldsymbol{Q} \sim W\left(\boldsymbol{\Sigma},\ n_2 - p\right)$ and thus $(-q_{12}/q_{11},\ q_{11}) \sim NGa\left[\lambda_{12},\ \lambda_{22},\ (n_2 - p)/2,\ \lambda_{11}/2\right]$ (Giri, 2003). If we write $\tilde{\lambda}_{12} = -a \cdot q_{12}/q_{11}$ we get $E\left(\tilde{\lambda}_{12}\right) = a \cdot \lambda_{12}$ and $var\left(\tilde{\lambda}_{12}\right) = a^2 \cdot \lambda_{11}/[(n_2 - p - 2)\,\lambda_{22}]$. Since we want to minimize the variance of the unbiased estimator for $\boldsymbol{\beta}_2$ we want to maximize

$$\begin{aligned} E_{\tilde{\lambda}_{12}}\left[f\left(\boldsymbol{\Lambda}, \tilde{\lambda}_{12}\right)\right] =& \lambda_{11}^{-1}\left\{2\lambda_{12}E\left(\tilde{\lambda}_{12}\right) - \left[E\left(\tilde{\lambda}_{12}\right)\right]^2 - var\left(\tilde{\lambda}_{12}\right)\right\}; \\ =& a\left(2\lambda_{12}^2/\lambda_{11} - a\left\{\lambda_{12}^2/\lambda_{11} + 1/[(n_2 - p - 2)\,\lambda_{22}]\right\}\right); \end{aligned}$$

with respect to a. If $\boldsymbol{\Lambda}$ was known, we see that this expectation is maximized for

$$a_{Oracle} = \left[(n_2 - p - 2)\,\lambda_{12}^2\lambda_{22}\right] / \left[(n_2 - p - 2)\,\lambda_{12}^2\lambda_{22} + \lambda_{11}\right];$$

Furthermore we see that if $0 \leq a \leq 2a_{Oracle}$, the variance of the unbiased estimator $\hat{\boldsymbol{\beta}}_{2\tilde{\lambda}_{12}}$ is smaller than the variance of the standard OLS estimator, $\hat{\boldsymbol{\beta}}_{22}$.

The fractions $n_i/(\alpha_i + n_i)$, $i = 1, 2$, might be seen as James–Stein shrinkage factors, corresponding to $k_{Oracle\,i}$ in Gangsei et al. (2016). Thus the estimator $\tilde{\boldsymbol{\beta}}_{2\,EB}$ might be seen as a double shrinkage estimator, where α_i, $i = 1, 2$, represents the prior population size for shrinking the estimator for $\boldsymbol{\beta}_i$ towards zero, and γ_2 is a shrinkage factor for the estimator of λ_{12}. Our hypothesis, in line with the well known principles of empirical Bayes strategies, is that tuning the values of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ by maximizing model evidence will be a suitable strategy for setting these parameters. We tested this assumption by a comprehensive simulation study.

5 Simulation study

5.1 Simulation design

To simulate data, we set up a simulation study using the software "R" (R Core Team, 2014) and an extension of the package "Simrel" (Sæbø, 2015; Sæbø et al., 2015), capable of producing a bivariate response variable. In the package, population values for $\mathbf{\Sigma}$, p and $\boldsymbol{R^2} = (R_1^2,\ R_2^2)^T$, i.e. the coefficients of determination, are possible inputs. The package also offers the opportunity to simulate extra predictor variables with no relevance to the response variable.

We used $19 \times 4 \times 3 \times 10 = 2280$ different combinations for the population parameters $\mathbf{\Sigma}$, $\boldsymbol{R^2}$, n_1 and n_2. The 19 settings of $\mathbf{\Sigma}$ were varied over $\rho = \sigma_{12}/\sqrt{\sigma_{11}\sigma_{22}} = -0.9, -0.8, \ldots, 0.9$. The four combinations $(0.75,\ 0.75)^T$, $(0.75,\ 0.25)^T$, $(0.25,\ 0.75)^T$ and $(0.5,\ 0.5)^T$ were used for $\boldsymbol{R^2}$. Finally we set n_1 and n_2 so that $n_2 = 6$, 7, 9, 10, 12, 14, 17, 20, 24, 28 was crossed over $n_1 = n_2 + 5, 35, 100$. Finally we set $p = 3$.

For each combination we used 5 different estimators for $\boldsymbol{\beta}_2$. (i) $\hat{\boldsymbol{\beta}}_{22}$, i.e. the ordinary OLS estimator, the two unbiased estimators (ii) $\hat{\boldsymbol{\beta}}_{2\tilde{\lambda}}$ and (iii) $\hat{\boldsymbol{\beta}}_{2\hat{\lambda}}$, (iv) $\tilde{\boldsymbol{\beta}}_{2\,EB}$, i.e. the empirical Bayes estimator and finally (v) $\tilde{\boldsymbol{\beta}}_2 = [n_2/(n_2+\alpha_2)]\,\hat{\boldsymbol{\beta}}_{2\tilde{\lambda}}$. $\tilde{\boldsymbol{\beta}}_2$ is not an Bayes estimator for the prior in question. However, unlike the real empirical Bayes estimator, $\tilde{\boldsymbol{\beta}}_{2\,EB}$, this estimator is coherent with the justification for empirical Bayes as the "adjustment part" is proportional to $\left(\hat{\boldsymbol{\beta}}_{21} - \hat{\boldsymbol{\beta}}_{21}\right)$, not to $\left\{[n_2/(n_2+\alpha_2)]\,\hat{\boldsymbol{\beta}}_{21} - \ [n_1/(n_1+\alpha_1)]\,\hat{\boldsymbol{\beta}}_{11}\right\}$.

It is worth noting that under the assumption of independent, identically normal distributed predictor variables, the expected risk functions for prediction error, i.e. $E_{\boldsymbol{x}}\left(R_{\hat{y}2N}\right)$, are known functions for $\hat{\boldsymbol{\beta}}_{22}$ and $\hat{\boldsymbol{\beta}}_{2\hat{\lambda}_{12}}$ (Gangsei et al., 2016). For the remaining estimators, i.e. the empirical Bayes–based estimators $\hat{\boldsymbol{\beta}}_{2\,EB}$ and $\tilde{\boldsymbol{\beta}}_{2\tilde{\lambda}_{12}}$, we lack closed form solutions for the expected prediction risks, and thus the primary aim of the simulation study is to evaluate these risks.

A total of 5×10^4 observations of $\boldsymbol{Y}$ and $\boldsymbol{X}$ were simulated via "Simrel" for each combination of model parameters. For 5000 repeats, we draw training sets (i.e. used to estimate model parameters) of size n_1 and test sets of size 1000 randomly from the 5×10^4 observations of $\boldsymbol{Y}$ and $\boldsymbol{X}$. The estimated risk for each of the 5000 repetitions was calculated as

$$\hat{R}_{\hat{y}xx\ j} = (1/1000)\sum_{i=1}^{1000}\left(\hat{y}_{i\ xx} - y_i\right)^2,\ \ j = 1, \ldots, 5000;$$

where the subindex $_{xx}$ indicates the estimator in question, i.e. $_{OLS}$, $_{EB}$ etc. The vectors of

$\boldsymbol{R}^2$	$(0.75, 0.75)$		$(0.75, 0.25)$		$(0.25, 0.75)$		$(0.5, 0.5)$	
	Rank	%	Rank	%	Rank	%	Rank	%
$\underline{n_1 = n_2 + 5}$								
$\hat{\boldsymbol{\beta}}_{22}$	3.7	-	4.2	-	2.9	-	3.4	-
$\tilde{\boldsymbol{\beta}}_{2\,EB}$	3.9	99	2.2	95	4.9	105	4.3	100
$\tilde{\boldsymbol{\beta}}_2$	2.1	98	2.1	95	2.3	98	2.2	97
$\hat{\boldsymbol{\beta}}_{2\hat{\lambda}}$	3.3	101	3.1	101	3.1	101	2.7	100
$\hat{\boldsymbol{\beta}}_{2\tilde{\lambda}}$	2.0	98	3.4	98	1.9	98	2.4	98
$\underline{n_1 = 35}$								
$\hat{\boldsymbol{\beta}}_{22}$	3.7	-	4.2	-	3.3	-	3.8	-
$\tilde{\boldsymbol{\beta}}_{2\,EB}$	2.9	97	1.8	94	4.3	99	3.2	96
$\tilde{\boldsymbol{\beta}}_2$	2.7	97	2.6	94	2.2	97	2.7	96
$\hat{\boldsymbol{\beta}}_{2\hat{\lambda}}$	3.6	100	3.0	100	3.4	100	2.8	100
$\hat{\boldsymbol{\beta}}_{2\tilde{\lambda}}$	2.2	96	3.5	97	1.8	97	2.5	96
$\underline{n_1 = 100}$								
$\hat{\boldsymbol{\beta}}_{22}$	3.7	-	4.2	-	3.6	-	3.8	-
$\tilde{\boldsymbol{\beta}}_{2\,EB}$	2.3	95	1.7	93	3.6	96	2.3	95
$\tilde{\boldsymbol{\beta}}_2$	3.3	95	2.8	93	2.5	96	3.2	95
$\hat{\boldsymbol{\beta}}_{2\hat{\lambda}}$	3.5	99	2.8	99	3.7	100	2.8	102
$\hat{\boldsymbol{\beta}}_{2\tilde{\lambda}}$	2.3	95	3.5	96	1.6	96	2.8	95

Table 1: Summary statistics for comparing the 5 estimators $\hat{\boldsymbol{\beta}}_{22}$, $\tilde{\boldsymbol{\beta}}_{2\,EB}$, $\tilde{\boldsymbol{\beta}}_2$, $\hat{\boldsymbol{\beta}}_{2\tilde{\lambda}}$ and $\hat{\boldsymbol{\beta}}_{2\hat{\lambda}}$. Each row represents one of the estimators. The upper 5 rows represent results when $n_1 = n_2 + 5$, the next 5 rows represent $n_1 = 35$ and the last 5 rows represent $n_1 = 100$. Double columns represent different combinations for $\boldsymbol{R}^2$ as presented in the heading. For every combination of $\boldsymbol{R}^2$, the leftmost column shows the average rank (1 - 5), i.e. estimators were ranked according to prediction precision, where 1 is best, and the rightmost column shows the average prediction error as % of prediction error using the OLS estimator. All averages are calculated as averages of the 10 combinations of n_2 crossed over the 19 combinations of ρ.

length 5000 of estimated risks were used to estimate the expected risks, denoted $\hat{E}_{\boldsymbol{x}}\left(R_{\hat{y}xx}\right)$, as means of these vectors. Furthermore we calculated 2.5% -, 97.5% quantiles and medians based on the same vectors. We calculated means and quantiles for the elements of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ by the same principle.

5.2 Simulation results

For the "oracle parameters" evaluated in Gangsei et al. (2016) we showed that there was a hierarchy between the estimators such that some of the estimators were uniformly better than others in the sense of having lower expected prediction error. It was also shown that the relationship between expected prediction error based on the estimator, $\hat{\boldsymbol{\beta}}_{2\hat{\lambda}}$, i.e. $E_{\boldsymbol{x}}\left(R_{\hat{y}\hat{\lambda}}\right)$, and the expected prediction error based on the ordinary OLS estimator, i.e. $E_{\boldsymbol{x}}\left(R_{\hat{y}\text{OLS}}\right)$, vary in the sense that for some combinations of model parameters, $\hat{\boldsymbol{\beta}}_{2\hat{\lambda}}$ is preferable over $\hat{\boldsymbol{\beta}}_{22}$,

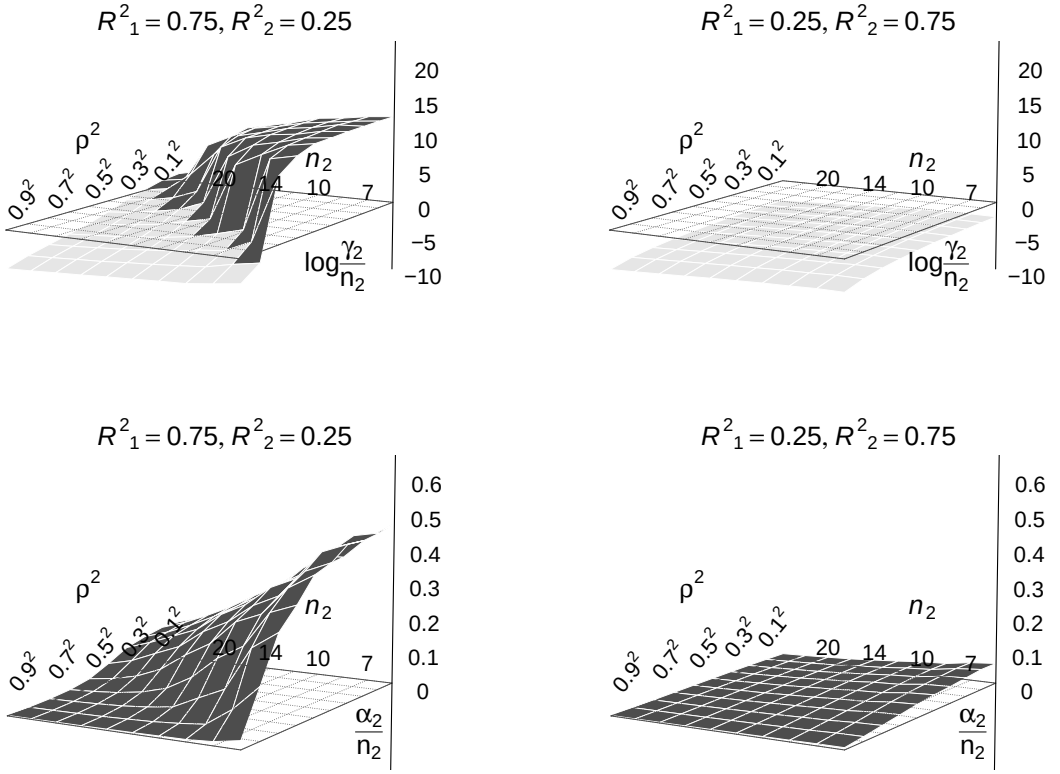


Figure 1: The upper panels show median values for $\log(\gamma_2/n_2)$ crossed over different combinations for ρ^2 and n_2. The lower panels show median values for (α_2/n_2) over the same combinations of ρ^2 and n_2. n_1 is 100 for all panels. For left panels $\boldsymbol{R}^2 = (0.75, 0.25)$ and for right panels $\boldsymbol{R}^2 = (0.25, 0.75)$.

typically when ρ^2 is "large", and for other model-parameter combinations $\hat{\boldsymbol{\beta}}_{22}$ is preferable over $\hat{\boldsymbol{\beta}}_{2\hat{\lambda}}$. This relationship is illustrated in the right side panels of Figure 3.

When using the simulation setup, we have a total of $4(R^2)\times 19(\rho)\times 10(n_2)\times 3(n_1) = 4560$ combinations by which we can assess the expected prediction error for different estimators against each other. The main results are summarized in Table 1. We observe that the empirical Bayes estimators perform well. Furthermore, the biased estimators perform worse when R_2^2 is low and R_1^2 is high than for the opposite situation. This is especially evident in combination with small n_1.

Table 1 shows that when we average over ρ and n_2, then $\hat{\boldsymbol{\beta}}_{2\tilde{\lambda}}$ outperforms $\hat{\boldsymbol{\beta}}_{2\hat{\lambda}}$ for all (simulated) combinations of $\boldsymbol{R}^2$ and n_1. The natural interpretation of this result is that maximizing model evidence is a powerful tool for setting γ_2 at suitable values. Figure 1 shows that γ_2 and α_2 respond in the expected and desired way when ρ, n_2 and $\boldsymbol{R}^2$ vary. The shrinkage increases as R_2^2 decreases, n_2 decreases and ρ^2 decreases. We also observe that when R^2 and n_2 are kept fixed, α_2 decreases as ρ^2 increases.

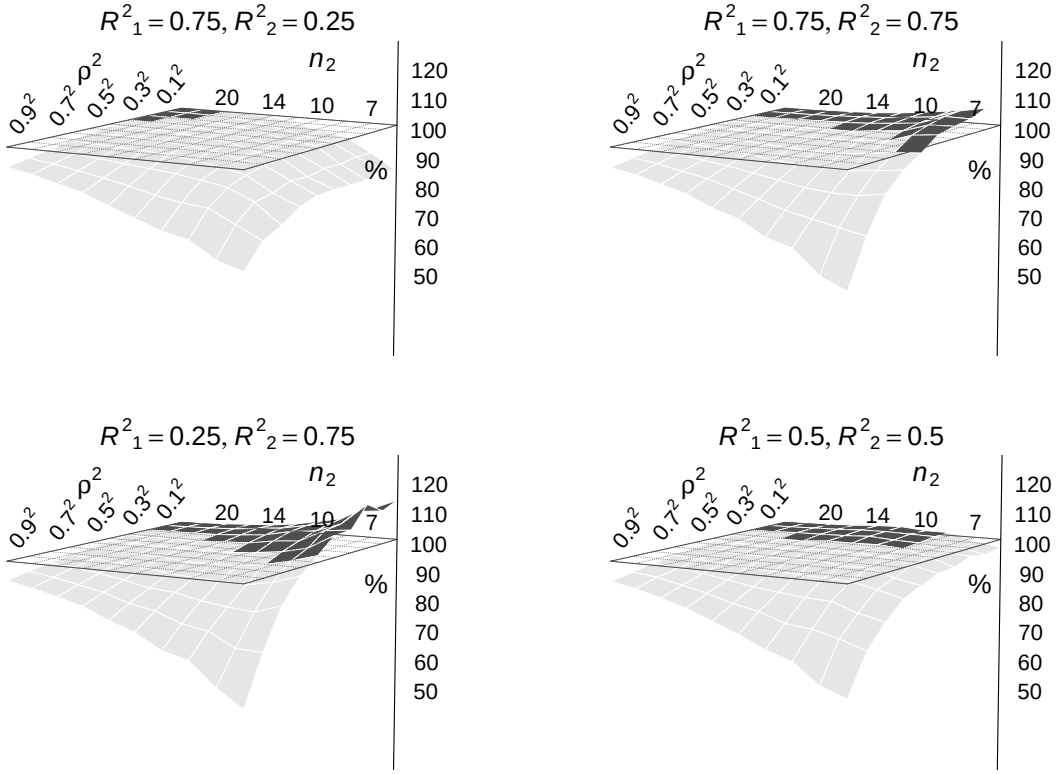


Figure 2: Simulated ratio between expected prediction error using the empirical Bayes estimator and the ordinary OLS estimator, i.e. $\hat{E}_{\boldsymbol{x}}\left(R_{\hat{y}\ EB}\right)/\hat{E}_{\boldsymbol{x}}\left(R_{\hat{y}\text{OLS}}\right)$ for different combinations of $\boldsymbol{R^2}$. From upper left to lower right panel; $\boldsymbol{R^2} = (0.75,\ 0.25)$, $\boldsymbol{R^2} = (0.75,\ 0.75)$, $\boldsymbol{R^2} = (0.25,\ 0.75)$ and $\boldsymbol{R^2} = (0.5,\ 0.5)$. For all panels $p = 3$ and $n_1 = 100$. In each panel, results for $\rho^2 = 0.9^2, 0.8^2, \ldots, 0^2$ and $n_2 = 6, 7, 9, 10, 12, 14, 17, 20, 24, 28$ are shown. Dark areas represent combinations of n_2 and ρ^2 where the OLS estimator has a lower expected prediction error.

6 Discussion

Figures 2 and 3 are examples meant to illustrate some general patterns. We have examined similar figures for all combinations of simulated model parameters. Figure 2 illustrates how the empirical Bayes estimator outperforms the ordinary OLS estimator for most combinations of model parameters. Just as expected, the gain is high when R_1^2 is large compared with R_2^2, and when ρ^2 is high. The empirical Bayes estimator even outperforms the OLS estimator for some combinations of R_1^2 and R_2^2 even if $\rho^2 = 0$. This is due to the James–Stein shrinkage effect. As seen by Figure 3 the unbiased estimators never outperforms the OLS estimator for $\rho^2 = 0$.

Moreover, we observed that the performance of the estimator $\hat{\boldsymbol{\beta}}_{2\hat{\lambda}_{12}}$ is extremely variable for different values of ρ^2. For ρ^2 values close to 1, $\hat{\boldsymbol{\beta}}_{2\hat{\lambda}_{12}}$ mostly outperforms the other estimators, but in general the differences in relation to the empirical Bayes based estimators

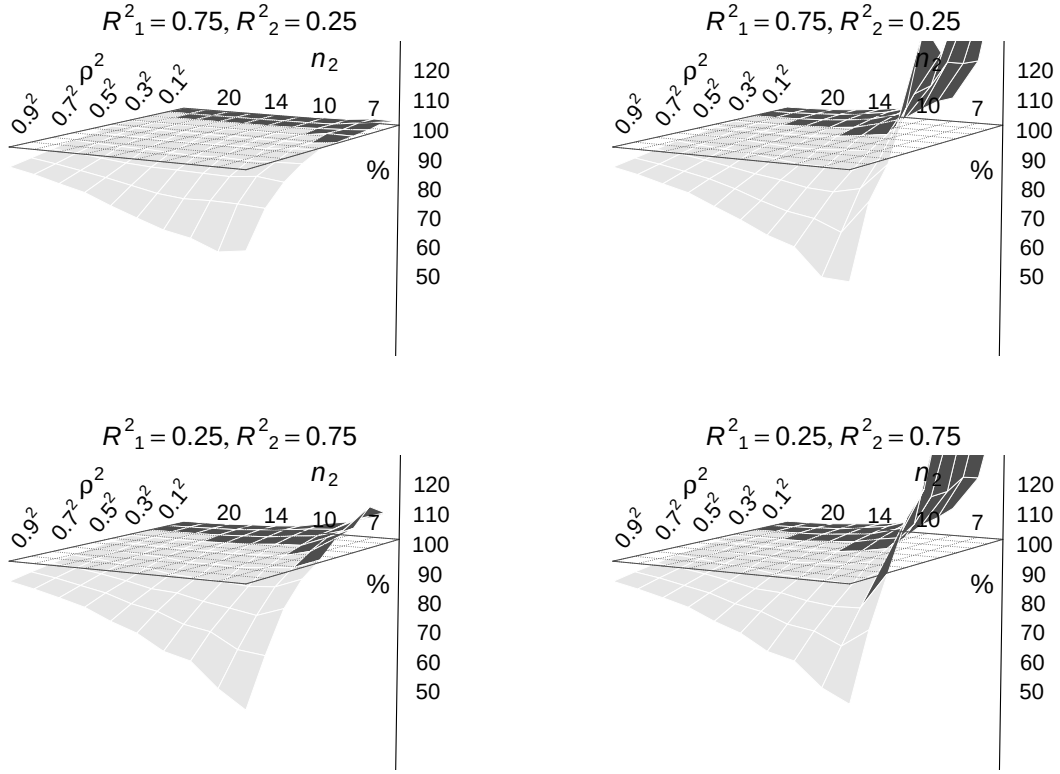


Figure 3: Simulated ratio between expected prediction error using the estimators $\hat{\boldsymbol{\beta}}_{2\tilde{\lambda}}$ (left panels), and $\hat{\boldsymbol{\beta}}_{2\hat{\lambda}}$ (right panels), and the ordinary OLS estimator, i.e. $\hat{E}_{\boldsymbol{x}}\left(R_{\hat{y}\hat{\lambda}}\right)/\hat{E}_{\boldsymbol{x}}\left(R_{\hat{y}\text{OLS}}\right)$ and $\hat{E}_{\boldsymbol{x}}\left(R_{\hat{y}\tilde{\lambda}}\right)/\hat{E}_{\boldsymbol{x}}\left(R_{\hat{y}\text{OLS}}\right)$. Two different combinations of $\boldsymbol{R^2}$, $\boldsymbol{R^2} = (0.75,\ 0.25)$ (upper panels) and $\boldsymbol{R^2} = (0.25,\ 0.75)$ (lower panels) are shown. For all panels $p = 3$ and $n_1 = 100$. In each panel, the results for $\rho^2 = 0.9^2, 0.8^2, \ldots, 0^2$ and $n_2 = 6, 7, 9, 10, 12, 14, 17, 20, 24, 28$ are shown.

are small. However, for smaller values of ρ^2, the estimator $\hat{\boldsymbol{\beta}}_{2\hat{\lambda}_{12}}$ often performs very poorly as shown in Figure (3).

We do not claim that the form of the prior distribution that we choose is the best. It might be possible for this prior, or maybe for a somewhat different prior distribution, based on similar principles, to find analytical solutions for the parameter values of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$ (or similar parameters) that maximize the model evidence. Then it might also be possible to find analytical solutions for expected prediction error.

When ρ is 0, general theory tells us that the OLS estimator is the best unbiased estimator in the sense of having lowest variance. Thus, it is not surprising that our alternative candidates, at least the unbiased candidates, perform worse than the OLS estimate for small values of ρ.

References

Box, G. and Tiao, G. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley series in behavioral science: quantitative methods. Addison-Wesley Pub. Co.

DeGroot, M. H. (2005). *Optimal Statistical Decisions*, volume 82. John Wiley & Sons.

Efron, B. and Morris, C. (1971). Limiting the Risk of Bayes and Empirical Bayes Estimators-Part I: The Bayes Case. *J. Amer. Statist. Assoc.*, 66(336):807–815.

Efron, B. and Morris, C. (1972a). Empirical Bayes on vector observations: An extension of Stein's method. *Biometrika*, 59(2):335–347.

Efron, B. and Morris, C. (1972b). Limiting the Risk of Bayes and Empirical Bayes Estimators-Part II: The Empirical Bayes Case. *J. Amer. Statist. Assoc.*, 67(337):130–139.

Efron, B. and Morris, C. (1973). Stein's Estimation Rule and its Competitors-An Empirical Bayes Approach. *J. Amer. Statist. Assoc.*, 68(341):117–130.

Efron, B. and Morris, C. (1975). Data Analysis Using Stein's Estimator and its Generalizations. *J. Amer. Statist. Assoc.*, 70(350):311–319.

Efron, B. and Morris, C. (1976). Families of Minimax Estimators of the Mean of a Multivariate Normal Distribution. *Ann. Statist.*, 4:11–21.

Gangsei, L. E., Almøy, T., and Sæbø, S. (2016). Theoretical evaluation of prediction error in linear regression with a bivariate response variable containing missing data. Submitted manuscript to Communications in Statistics – Theory and Methods.

Giri, N. C. (2003). *Multivariate Statistical Analysis: Revised and Expanded*, volume 171. CRC Press.

James, W. and Stein, C. (1961). Estimation with quadratic loss. In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 361–379.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, 90(430):773–795.

Minka, T. (2000). Bayesian linear regression. Technical report, Microsoft Research Cambridge.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sæbø, S. (2015). *simrel: Linear Model Data Simulation and Design of Computer Experiments*. R package version 1.1-0.

Sæbø, S., Almøy, T., and Helland, I. S. (2015). simrel-A versatile tool for linear model data simulation based on the concept of a relevant subspace and relevant predictors. *Chemometrics and Intelligent Laboratory Systems*.

Upton, G. and Cook, I. (2014). *A Dictionary of Statistics 3e*. Oxford university press.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243.

A Appendix

A.1 Posterior means and variances

$$E(\lambda_{11}) = (\zeta_1 + n_1)/\upsilon_3,\ \ var(\lambda_{11}) = 2(\zeta_1 + n_1)/\upsilon_3^2;$$

$$E(\lambda_{12}) = -\upsilon_{12}/\upsilon_{11},\ \ var(\lambda_{12}) = |\boldsymbol{\Upsilon}|/\left[\upsilon_{11}^2(\zeta_2 + n_2 - 2)\right];$$

$$E(\lambda_{22}) = (\zeta_2 + n_2)/\left[\upsilon_{22} - \left(\upsilon_{12}^2/\upsilon_{11}\right)\right],\ \ var(\lambda_{22}) = 2\left\{(\zeta_2 + n_2)/\left[\upsilon_{22} - \left(\upsilon_{12}^2/\upsilon_{11}\right)\right]^2\right\};$$

$$E(\boldsymbol{\beta}_1) = \left(\boldsymbol{\Psi_1} + \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1}\left(\boldsymbol{X}_1^T\boldsymbol{y}_1 + \boldsymbol{\Psi}_1\boldsymbol{\beta}_{01}\right),\ \ var(\boldsymbol{\beta}_1) = \left[(\zeta_1 + n_1)/\upsilon_3\right]\left(\boldsymbol{\Psi_1} + \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1};$$

$$E(\boldsymbol{\beta}_2) = \left(\boldsymbol{\Psi_2} + \boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2^T\boldsymbol{y}_{22} + \boldsymbol{\Psi}_2\boldsymbol{\beta}_{02}\right);$$
$$- (\upsilon_{12}/\upsilon_{11})\left[\left(\boldsymbol{\Psi_2} + \boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1}\left(\boldsymbol{X}_2^T\boldsymbol{y}_{21} + \boldsymbol{\Psi}_2\boldsymbol{\beta}_{01}\right) - E(\boldsymbol{\beta}_1)\right];$$

$$var(\boldsymbol{\beta}_2) = \left\{|\boldsymbol{\Upsilon}|/\left[(\zeta_2 + n_2 - 2)\,\upsilon_{11}^2\right]\right\}\left(\hat{\boldsymbol{\beta}}_{21} - \hat{\boldsymbol{\beta}}_1\right)\left(\hat{\boldsymbol{\beta}}_{21} - \hat{\boldsymbol{\beta}}_1\right)^T;$$
$$+ \left\{|\boldsymbol{\Upsilon}|/\left[(\zeta_2 + n_2 - 2)\,\upsilon_{11}\right]\right\}\left(\boldsymbol{\Psi_2} + \boldsymbol{X}_2^T\boldsymbol{X}_2\right)^{-1};$$
$$+ \left\{\left[\upsilon_3|\boldsymbol{\Upsilon}| + (\zeta_2 + n_2 - 2)\,\upsilon_{12}^2\upsilon_3\right]/\left[(\zeta_1 + n_1 - 2)(\zeta_2 + n_2 - 2)\,\upsilon_{11}^2\right]\right\}\left(\boldsymbol{\Psi_1} + \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1};$$

$$cov(\boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2) = \left\{(\upsilon_{12}\upsilon_3)/\left[\upsilon_{11}(\zeta_1 + n_1 - 2)\right]\right\}\left(\boldsymbol{\Psi_1} + \boldsymbol{X}_1^T\boldsymbol{X}_1\right)^{-1};$$

A.2 Equations for empirical Bayes hyperparameters

The empirical Bayes hyperparameters were set by

$$\begin{aligned}(\alpha_1,\,\gamma_1) = argmax_{(\alpha_1,\gamma_1)}\, &\{\log\Gamma\left[(n_1+\gamma_1)/2\right] - \log\Gamma\left(\gamma_1/2\right) + \\ &(\gamma_1/2)\log(\gamma_1) - (n_1+\gamma_1-1)/2\cdot\log(\upsilon_3) + \\ &(p/2)\log\left[\alpha_1/(n_1+\alpha_1)\right]\}\,; \\ (\alpha_2,\,\gamma_2) = argmax_{(\alpha_2,\gamma_2)}\, &\{\log\Gamma\left[(n_2+\gamma_2)/2\right] - \log\Gamma\left(\gamma_2/2\right) + \\ &[(\gamma_2-1)/2]\cdot\log(\gamma_2) - [(n_2+\gamma_2)/2]\log(|\boldsymbol{\Upsilon}|) + \\ &[(n_2+\gamma_2-1)/2]\log\left(\upsilon_{11}\right) + (p/2)\log\left[\alpha_2/(n_2+\alpha_2)\right]\}\,;\end{aligned}$$

Note that $\boldsymbol{\Upsilon}$ and υ_3 both are functions of $\boldsymbol{\alpha}$ and $\boldsymbol{\gamma}$.

A.3 Proof Theorem 1

By using some tedious, but straightforward matrix algebra, it is possible to show that the likelihood function, denoted $\pi\left(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Lambda}\right)$, might be written as

$$\begin{aligned}\pi\left(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Lambda}\right) = &(2\pi)^{-(n_1+n_2)/2}\,\lambda_{11}^{n_1/2}\lambda_{22}^{n_2/2} \\ &\exp\Big\{-(1/2)\Big[\lambda_{22}\left(q_{22}-q_{12}^2/q_{11}\right) + \lambda_{22}q_{11}\left(\lambda_{12}--q_{12}/q_{11}\right)^2 \\ &+\lambda_{11}q_3 + \left(\boldsymbol{\beta}_v-\hat{\boldsymbol{\beta}}_v\right)^T\boldsymbol{\Omega}_X^{-1}\left(\boldsymbol{\beta}_v-\hat{\boldsymbol{\beta}}_v\right)\Big]\Big\}\,;\end{aligned}$$

Furthermore, it might be shown that the product of the likelihood, $\pi\left(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Lambda}\right)$, and prior distribution, $\pi\left(\boldsymbol{\beta}, \boldsymbol{\Lambda}\right)$, might be written as

$$\pi\left(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Lambda}\right)\times\pi\left(\boldsymbol{\beta}, \boldsymbol{\Lambda}\right) = g_0\left(\boldsymbol{\eta}\right)\times f_1\left(\lambda_{11}\right)\times f_2\left(\lambda_{22}\right)\times f_3\left(\lambda_{12},\lambda_{22}\right)\times f_4\left(\boldsymbol{\beta}_v,\boldsymbol{\Lambda}\right);$$

, where

$$\begin{aligned}f_1\left(\lambda_{11}\right) =& \lambda_{11}^{[(n_1+\zeta_1)/2-1]}\exp\left[-\left(1/2\right)\upsilon_3\lambda_{11}\right]; \\ f_2\left(\lambda_{22}\right) =& \lambda_{22}^{[(n_2+\zeta_2)/2-1]}\exp\left[-\left(1/2\right)\left(\upsilon_{22}-\upsilon_{12}^2/\upsilon_{11}\right)\lambda_{22}\right]; \\ f_3\left(\lambda_{12},\lambda_{22}\right) =& \left(\lambda_{22}\upsilon_{11}\right)^{1/2}\exp\left[-\left(1/2\right)\lambda_{22}\upsilon_{11}\left(\lambda_{12}-\upsilon_{12}/\upsilon_{11}\right)^2\right]; \\ f_4\left(\boldsymbol{\beta}_v,\boldsymbol{\Lambda}\right) =& \left|\tilde{\boldsymbol{\Omega}}\right|^{-1/2}\exp\left[\left(1/2\right)\cdot\left(\boldsymbol{\beta}_v-\tilde{\boldsymbol{\beta}}_v\right)^T\tilde{\boldsymbol{\Omega}}^{-1}\left(\boldsymbol{\beta}_v-\tilde{\boldsymbol{\beta}}_v\right)\right];\end{aligned}$$

, and where

$$g_0(\boldsymbol{\eta}) = \prod_{i=1}^{2} \left[\frac{(2\pi)^{-(n_i+p+0.5)/2}}{2^{\zeta_i/2}\Gamma(\zeta_i/2)} \cdot \frac{|\boldsymbol{\Psi}_i|^{1/2}}{|\boldsymbol{\Psi}_i + \boldsymbol{X}_i^T \boldsymbol{X}_i|^{1/2}} \right] \cdot \frac{|\boldsymbol{\Phi}|^{\zeta_2/2}}{\phi_3^{-\zeta_1/2} \phi_{11}^{1/2} \upsilon_{11}^{1/2}};$$

is a constant independent of $\boldsymbol{\Lambda}$ and $\boldsymbol{\beta}$.

From the form of the functions $f_i(\cdot)$ it is clear that the posterior must have the distribution as given by Eq. (2).

A.4 Proof Lemma 1

Due to the form of the posterior distribution, $\pi(\boldsymbol{\beta}, \boldsymbol{\Lambda} \mid \boldsymbol{Y})$, it might be written as

$$\pi(\boldsymbol{\beta}, \boldsymbol{\Lambda}|\boldsymbol{Y}) = \prod_{i=1}^{4} g_i(\boldsymbol{\eta}) f_i(\cdot);$$

, where the functions $g_i(\boldsymbol{\eta})$ are independent of $\boldsymbol{\Lambda}$ and $\boldsymbol{\beta}$ and given by

$$\begin{aligned} g_1(\boldsymbol{\eta}) &= \frac{(\upsilon_3/2)^{(n_1+\zeta_1)/2}}{\Gamma[(n_1+\zeta_1)/2]}; \quad g_2(\boldsymbol{\eta}) = \frac{[(\upsilon_{22} - \upsilon_{12}^2/\upsilon_{11})/2]^{(n_2+\zeta_2)/2}}{\Gamma[(n_2+\zeta_2)/2]}; \\ g_3(\boldsymbol{\eta}) &= (1/2\pi)^{1/2}; \qquad g_4(\boldsymbol{\eta}) = (1/2\pi)^p; \end{aligned}$$

A slight transformation of Bayes theorem reveals a convenient formula for the model evidence as the quotient between "prior times likelihood" and "posterior", i.e.

$$\pi(\boldsymbol{Y}) = \frac{\pi(\boldsymbol{Y} \mid \boldsymbol{\beta}, \boldsymbol{\Lambda}) \times \pi(\boldsymbol{\beta}, \boldsymbol{\Lambda})}{\pi(\boldsymbol{\beta}, \boldsymbol{\Lambda}|\boldsymbol{Y})} = \frac{g_0(\boldsymbol{\eta})}{\prod_{i=1}^{4} g_i(\boldsymbol{\eta})};$$

Evaluation of this quotient leads to the formula shown in Eq. (3).

A.5 Distributions and notations

The following notations are used for different well known distributions:

- $x \sim N(\mu, \sigma^2)$, for the normal distribution where μ denotes expectation and σ^2 denotes the variance.
- $x \sim Ga(\alpha, \beta)$, for the gamma distribution where α denotes shape parameter and β denotes the rate parameter.

- $(x_1,\ x_2) \sim\ NGa\,(\mu,\ \tau,\ \alpha,\ \beta)$. For the normal–gamma distribution. The normal–gamma distribution is a compound distribution where $x_2 \sim Ga\,(\alpha,\ \beta)$ and $x_1 \mid x_2 \sim N\,[\mu,\ 1/(\tau x_2)]$.

- $\boldsymbol{x} \sim\ N_m\,(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$, for the multivariate normal distribution, where $\boldsymbol{\mu}$ denotes the expectation vector and $\boldsymbol{\Sigma}$ denotes the covariance matrix.

- $\boldsymbol{X} \sim\ MN_{m,p}\,(\boldsymbol{\mu},\ \boldsymbol{\Sigma}_1,\ \boldsymbol{\Sigma}_2)$, for the matrix–normal distribution, where $\boldsymbol{X}$ is a matrix of size $m \times p$, $\boldsymbol{\mu}$ denotes the location matrix $(m \times p)$ and $\boldsymbol{\Sigma}_1$ $(m \times m)$ and $\boldsymbol{\Sigma}_2$ $(p \times p)$ denotes scale matrixes.

- $\boldsymbol{X} \sim\ W\,(\boldsymbol{\Phi},\ \nu)$, for the Wishart distribution, where $\boldsymbol{\Phi}$ denotes the scale matrix and ν the degrees of freedom.

- $(\boldsymbol{X}_1,\ \boldsymbol{X}_2) \sim\ MNW\,(\boldsymbol{\mu},\ \boldsymbol{\Sigma}_1,\ \boldsymbol{\Phi},\ \nu)$, for the matrix–normal–Wishart distribution. The matrix–normal–Wishart distribution is a compound distribution where $\boldsymbol{X}_2 \sim W\,(\boldsymbol{\Phi},\ \nu)$ and $\boldsymbol{X}_1 \mid \boldsymbol{X}_2 \sim\ MN\,(\boldsymbol{\mu},\ ,\boldsymbol{\Sigma}_1,\ \boldsymbol{X}_2)$.

Paper III

Prediction precision for lean meat percentage in Norwegian pig carcasses using "Hennessy grading probe 7". Evaluation of methods emphasized at exploiting additional information from Computed Tomography.

Lars Erik Gangsei[a,b], Jørgen Kongsro[c], Eli Vibeke Olsen[d], Morten Røe[a], Ole Alvseike[a], Solve Sæbø[b]

Abstract: The present study aims at improving the prediction of lean meat percentage (LMP) for pig carcasses based on on–line measurements from the slaughterhouses using the "Hennessy Grading Probe 7" (HGP7) and auxiliary information such as gender and breed. The prediction performance is evaluated using an empirical Bayes method capable of utilising information from a surrogate variable, i.e. LMP from computed tomography.

HGP7 measures thicknesses of fat and meat layers. The HGP7 measurements of subcutaneous fat, sirloin height and interior fat layer should be included as predictor variables together with gender. For efficiency at the slaughter–line gender might be omitted.

The empirical Bayes method improved prediction precision only marginally compared with the standard ordinary least squares method when applied to the full set of data. However, simulations show that the empirical Bayes method enables a considerable reduction of the data sample size without appreciable loss of prediction precision.

Keywords: Computed tomography; empirical Bayes; grading probe; lean meat percentage; pig–carcass; prediction precision; surrogate variable

a: Animalia, P.O. Box 396 – Økern, N–0513 Oslo, Norway

b: Norwegian University of Life Sciences (NMBU), Department of Chemistry, Biotechnology and Food Science, P.O. Box 5003, N–1432 Ås, Norway

c: Norsvin SA, P.O. Box 504, N–2304 Hamar, Norway

d: Danish Meat Research Institute (DMRI), Gregersensvej 9, DK–2630 Taastrup, Denmark

1 Introduction

The lean meat percentage (LMP) is used as the primary classification variable for the quality of pig carcasses in the European Union (EU). LMP is defined as the proportion of weight of lean meat to the total weight of the carcass (Commission of the European Communities, 2008). Measuring the LMP is typically done using ultrasound or optical probes. Norwegian slaughterhouses use the optical probe "Hennessey Grading Probe 7" (HGP7).

The data obtained from optical probes or ultrasound do not provide a direct measurement of LMP, but are used as predictor variables in a regression equation whose sole purpose is LMP–prediction. Due to rapid evolution in the pig–population, and to maintain a high public confidence in on–line predictions of LMP, the prediction equation for LMP is updated on regular basis. In Norway this is done approximately every fifth year, including the years 2008 and 2013.

The parameters in the regression equation are estimated using a training sample of pig–carcasses where LMP is measured by at least one independent reference method. Until January 1st 2009 manual dissection (Walstra and Merkus, 1995) was the only method approved as an official reference method by the EU–authorities. From this date LMP–estimates from computed tomography (CT) was approved as an official reference method for LMP (Commission of the European Communities, 2008). Throughout this text LMP–MD and LMP–CT will denote LMP obtained by manual dissection and CT respectively.

EU regulations imposes requirements to prediction precision for LMP stating that: "*Grading methods shall be authorised only if the root mean squared error of prediction (*RMSEP*), computed by a full cross–validation technique or by a test set validation on a representative sample of at least 60 carcases, is less than 2,5. In addition, any outliers shall be included in the calculation of* RMSEP" (Commission of the European Communities, 2008).

Most countries use an equation which predicts LMP at a "manual dissection scale". However, CT scanning is cheaper than manual dissection. Further the LMP–CT estimates are believed to be more consistent, they are possible to replicate and they have a very high correlation with LMP–MD (Judas et al., 2007; Vester-Christensen et al., 2009). Therefore many countries use CT–dissection as their main reference method (Daumas and Monziols, 2011), usually supported by manual dissection for a subset of the carcasses.

The choice of method for estimating regression parameters is a trade off between complexity and accuracy. We want a method obtaining regression parameters providing small prediction errors for the whole population of pig–carcasses. On the other hand we want the method to be simple, and preferably yielding easily interpretable results. In this study

we examine a totally data–driven empirical Bayes method (Gangsei et al., 2016) capable of utilizing information from the additional observations of LMP–CT to improve regression estimates for LMP–MD.

In addition to find a suitable regression equation for LMP–MD based on the predictor variables in question this study aim at examine the relationship between sample sizes of manual dissected and CT–scanned carcasses and prediction precision for LMP–MD.

The program *R* (R Core Team, 2014), including the package *bestglm* (McLeod and Xu, 2011), was used for statistical computing. The R–code used in this study can be provided upon request.

2 Material and Methods

2.1 Data

The data consist of CT scans of 465 Norwegian half pig–carcasses of which 229 were scanned in 2008 and 236 in 2013. The pigs were slaughtered at two different commercial Norwegian abattoirs. The CT scanning was performed by Danish Meat Research Institute (DMRI). The carcasses were prepared, scanned and weighed in line with the description in Vester-Christensen et al. (2009). Based on these CT–data DMRI predicted LMP–CT for all carcasses using the method described in Vester-Christensen et al. (2009).

The carcasses were stratified to 4×4 different classes based on weight and on–line slaughterhouse HGP7 measurements to ensure data sampling across the entire range of carcasses. The HGP7–probe penetrates the rind, and measures the thickness of subcutaneous fat (denoted *Fat*), sirloin (denoted *Meat*) and the interior fat layer (denoted *Totif*) under the peritoneum. The measurements are done at two specified locations, one behind the last rib and 8 cm from the spine midline, and the other 12 cm further forward on the carcass, 6 cm from the spine midline. For all carcasses on–line data from the HGP7, weight, gender (castrates or females) and breed were registered.

A sample of the CT–scanned carcasses were transferred to Animalia's pilot plant in Oslo where manual dissection was carried out as described by Walstra and Merkus (1995), and the LMP–MD was calculated using the formula according to Commission of the European Communities (2006). A total of 86 carcasses were manually dissected, of which 66 were dissected in 2008 and 20 in 2013.

The maximum number of primary predictor variables was seven, of which four were continuous variables, *Fat*, *Meat*, *Totif* (from HGP7–measurement) and *Weight*. Note that

Fat was the mean of the two fat–measurements, since they were highly correlated (ρ = .824), whereas the measurements for *Meat* and *Totif* were single measurements from the foremost measurement point. The correlations between the four continuous predictor variables were low to moderate, i.e. in the range -0.49 till 0.48. In accordance with Gangsei et al. (2016) continuous predictor variables, and the response variables, were centered prior to the analysis.

There were three factorial variables. *Year* had two levels, 2008 (n=229) and 2013 (n=236). The variable *Gender* had two levels, females (n=228) and castrates (n=237). *Breed* had three levels; Hampshire (LYHH) (n=69), Duroc (LYLD) (n=243) and Norhybrid (LYLY) (n=153), representing three different hybrids used in Norway. The maternal line for all three hybrids were crossings between Norwegian Landrace and Yorkshire (denoted LY). The paternal lines were pure Hampshire (HH), crossing between Norwegian Landrace and Duroc (LD) and crossing between Norwegian Landrace and Yorkshire (LY) respectively. In the appendix (A.1) summary statistics for LMP–MD, LMP–CT and the continuous predictor variables crossed over the factor variables *Year*, *Gender* and *Breed* are presented. Since no Norhybrids were manually dissected in 2013, the variables *Year* and *Breed* were confounded for estimates based solely on LMP–MD, but not for estimates based on LMP–CT.

2.2 Model

In this paper we use notations in line with Gangsei et al. (2016). LMP–MD is a vector of length $n_2 = 86$ denoted $\boldsymbol{y}_{MD}$ (corresponding to $\boldsymbol{y}_{22}$ in Gangsei et al. (2016)), and LMP–CT is a vector of length $n_1 = 465$, denoted $\boldsymbol{y}_{CT}$ (corresponding to $\boldsymbol{y}_{11}$). Thus, the $n_1 \times 2$ matrix $\boldsymbol{Y}_1$ contains a substantial number of unobserved LMP–MD's.

The matrix $\boldsymbol{X}_1$ is an $n_1 \times p$ matrix of predictor variables where the first column is the unit vector. The next $p-1$ columns are the predictor variables. In some parts of the analysis second–order interaction terms are included as extra predictor variables in $\boldsymbol{X}_1$.

The formal model is given in (1), where $\boldsymbol{x}_i$ and $\boldsymbol{y}_i$ denotes the ith row of $\boldsymbol{X}_1$ and $\boldsymbol{Y}_1$ respectively. Note the matrix–form of the regression–parameter ($\boldsymbol{\beta} = [\boldsymbol{\beta}_{CT} \ \boldsymbol{\beta}_{MD}]$) and that the error–terms for each observation might be correlated with covariance matrix $\boldsymbol{\Sigma}$.

$$\boldsymbol{y}_i^T \overset{i.i.d.}{\sim} N_2\left(\boldsymbol{\beta}^T \boldsymbol{x}_i^T, \ \boldsymbol{\Sigma}\right), \ \ i = 1, \ldots, n_1; \tag{1}$$

Gangsei et al. (2016) deals with this model in detail under an empirical Bayes inference. Empirical Bayes methods uses the data to fit the prior distribution, including the hyper-parametes, i.e. the parameters used in the prior distrubution. In a strict Bayesian sense

this is "cheating". However, in a number of situations the empirical Bayes strategies are beneficial, as the prior is just a function of data, but also as the resulting Bayesian (biased) parameter estimates often are better (in some senses) than their unbiased counterparts (Carlin and Louis, 2008). One such case is linear regression, where the state–of–the–art empirical Bayes strategies leads to shrinkage estimators highly related to ridge regression estimators, estimators that often outperforms the standard ordinary least squares (OLS) estimator when it comes to prediction precision. In Gangsei et al. (2016) it is shown how this "empirical Bayes machinery" might be used in a situation with a bivariate response variable with missing data.

The denominator in the iconic Bayes theorem is known as the model evidence or the marginal likelihood. It might be viewed as a normalizing constant that ensures that the posterior distribution integrates to one over the model parameters. After the advent of computer technology the method known as Marcov Chain Monte Carlo sampling enables Bayesian examination of even extremely complex models, without the need or possibility of calculating the model evidence (Gilks et al., 1996). However, in some situations, like in Gangsei et al. (2016), the model evidence have an analytical expression. Since the model evidence might be given an interpretation as "likelihood of data conditional on model and hyperparameters", it might be used for model selection (Kass and Raftery, 1995). The well–known "Bayes factor" for comparing two models is simply the ratio of their model evidences.

Posterior means are used as point–estimates for $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$, and 95% Credibility intervals are calculated by Monte Carlo sampling based on 10^4 simulations.

2.3 Model comparison

A total of 177 different combinations of predictor variables, denoted "models" in the following paragraphs, were examined. These models included all 127 ($= 2^7 - 1$) possible combinations involving at least one of the predictor variables. Further, 50 models involving interaction–terms including the four continuous primary variables *Fat*, *Meat*, *Totif* and *Weight* as extra predictors were tested. These models were screened using BIC (Schwarz, 1978) as selection criteria via the R–package *bestglm* (McLeod and Xu, 2011) for models using LMP–CT as a single response variable. Within these 50 screened models the final model selection was done by examination of model evidence (Kass and Raftery, 1995) and RMSEP (root–mean–squared–error of prediction).

Different models, and the difference between using the empirical Bayes method and

OLS based on the full observations, were compared by computing RMSEP via leave–one–out cross–validation. If not stated otherwise RMSEP was based on prediction results for manual dissection (LMP–MD).

To investigate the gain of choosing the empirical Bayes method utilizing the extra information from CT, over OLS, in situations with different combinations of sample sizes for LMP–MD and LMP–CT, a resampling study was conducted on the model with the favored combination of predictor variables; *Fat*, *Meat*, *Totif* and *Gender*. RMSEP was calculated based on regression parameter estimates from OLS and the empirical Bayes method using sets of data where $n_2 = 10, 15, 20, 30, \ldots, 80$ observations of LMP–MD were assumed known. The subsets were sampled randomly from the real data on two restrictions; the sample should contain at least 2 carcasses from each sex, and the overall average *Fat* value was to fall inside the range of *Fat*–values in the sample.

For every subsample of LMP–MD the corresponding set of LMP–CT values were assumed known. In addition $n_1 - n_2 = 0, 10, 15, 25, 50, 100, 200, 300$ extra observations of LMP–CT were assumed known, by random sampling from the real data. For all combinations of n_2 and n_1 RMSEP was calculated using OLS, and the empirical Bayes method. The calculations were based on a test set comprised by the $86 - n_2$ ($n_2 = 86$ corresponds to the full set of data) observations of LMP–MD discarded from the sub–sample used for parameter estimation. The process was repeated 500 times for every value of n_2, except for $n_2 = 80$ where the process was repeated 1000 times due to the small test set size for this value of n_2.

3 Results

3.1 Prediction precision and model selection

As part of the preliminary work for this paper, the data was also analyzed using two alternative methods; two–stage least squares regression (2SLS) (Wooldridge, 2012, chap.15) and as a random effect model. To use the random effects model we had to assume that $\sigma_{12} > 0$ and that $\sigma_{11} = \sigma_{22}$. Both of these methods utilize the full set of data. The unreported results from these methods differed negligible from the results obtained by the empirical Bayes method.

The model referred to as "the favored model", is the model where HGP7–*variables* and *Gender* were used as predictor variables. Among the models not including interaction–terms this model had the highest model evidence for hyperparameters; $\gamma_1 = 1.90$, $\gamma_2 = 1.22$,

Table 1: 95% Credible Intervals, and posterior median and mean values, for the elements of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ using the favored model.

$\boldsymbol{\beta}_{CT}$	2.5%	50%	97.5%	Mean
Intercept	66.06	66.31	66.56	66.31
Fat	–1.18	–1.11	–1.04	–1.11
Meat	0.07	0.10	0.13	0.10
Totif	0.06	0.11	0.16	0.11
Gender	0.19	0.56	0.94	0.56
$\boldsymbol{\beta}_{MD}$	2.5%	50%	97.5%	Mean
Intercept	60.23	60.51	60.77	60.51
Fat	–1.10	–1.03	–0.95	–1.03
Meat	0.07	0.10	0.13	0.10
Totif	0.08	0.13	0.18	0.13
Gender	0.31	0.71	1.12	0.71
$\boldsymbol{\Sigma}$	2.5%	50%	97.5%	Mean
σ_{11} (CT)	3.17	3.60	4.09	3.61
σ_{12}	2.01	2.41	2.93	2.42
σ_{22} (MD)	2.28	2.67	3.12	2.68

$\alpha_1 = 2.06$ and $\alpha_2 = 2.65$. The hyperparameters might be given an interpretation as prior population size, see (Gangsei et al., 2016) for further interpretation. Table 1 shows estimates for regression parameters using the favored model. When interaction terms are not included the favored model minimizes RMSEP when using both OLS and the empirical Bayes method. This model, and the model where *Gender* was omitted, are marked with arrows in Figure 1.

Models including *Fat* as a predictor variable shows a clear pattern of having much higher model evidence and lower values for RMSEP compared to models where *Fat* is excluded, c.f. Figure 1. Inclusion of *Meat* and *Totif* as predictor variables unambiguously improves the model additionally. Among the factor variables *Gender* seems to be the only variable improving the predictive precision for LMP–MD as it is the only factor variable in the favored model using model evidence as selection criteria. Further *Gender* is close to significant using OLS (p=0.053). RMSEP decreased from 1.69 till 1.67 (OLS) and 1.67 till 1.63 (empirical Bayes) when *Gender* were added as predictor variable.

Inclusion of interaction–terms as predictor variables in the model gave no substantial improvement for prediction precision. 13 models including interaction–terms had marginally larger model evidence and 12 models had smaller RMSEP than the favored model. The smallest RMSEP using interactions was 1.62 compared to 1.63 for the favored model.

LMP–MD and LMP–CT were highly correlated, $\rho = .968$ (n=86). The effect of using the empirical Bayes method might be viewed as a way of "borrowing strength" from CT–data for

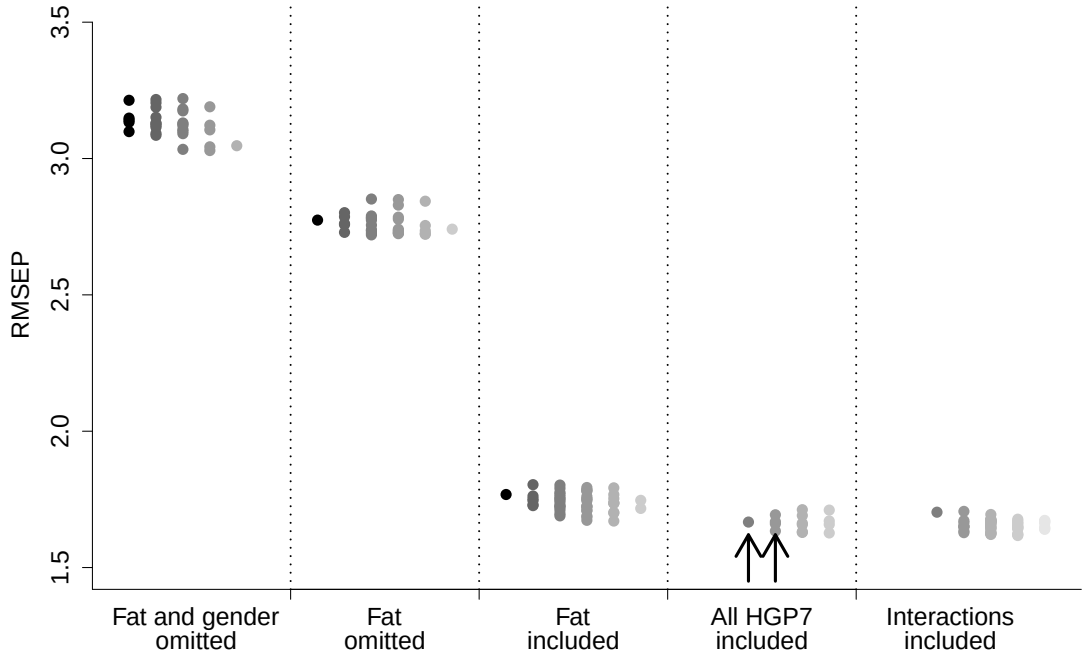


Figure 1: Relationship between number of predictor variables and RMSEP, using the empirical Bayes method, for all 127 possible model–combinations including of the primary predictor variables, and the 50 models including interaction–terms. Models omitting both *Fat* and gender as predictor variable are shown in the leftmost group. Models where *Gender*, but not *Fat* was included are shown in the second leftmost group. Models where *Fat* was included, but where *Meat* and/or *Totif* were omitted are shown in the middle group. The second rightmost group represents models where all HGP7–variables (*Fat*, *Meat* and *Totif*) were included, and finally the rightmost group represents the 50 models where interaction terms were included. Black color shows models with one predictor variable, then the gradual transition to the lightest grey represents 2,..., up till a maximum of 7 predictor variables. The two black arrows points at the favored model (right arrow) and the model where *Gender* is omitted (left arrow).

estimating the parameters associated with manual dissection. The high correlation between LMP–MD and LMP–CT is exploited and increases the effective test–sample size. Thereby the variance of prediction parameter estimates are reduced and prediction precision is increased.

Table 1 shows 95% Credible Intervals for parameters $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ for the favored model. Notice that the posterior distributions for the different elements of both $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ in general depend on each other. This is analogous to the situation in OLS where $cov\,\hat{\boldsymbol{\beta}} = \sigma^2 \left(\boldsymbol{X}^t \boldsymbol{X}\right)^{-1}$.

The parameter $\sigma_{22} = \sigma_{MD}$ has a natural interpretation as the expected variance for the prediction error for LMP–MD conditional on known $\boldsymbol{\beta}_{MD}$ and known values for the predictor variables ($\boldsymbol{X}$). Consequently, for the favored model, the lower limit for expected RMSEP was estimated to 1.63, i.e. the square root of posterior mean for σ_{22} (see Table 1). This is equal to the observed RMSEP. However, do note that the empirical Bayes method is a method

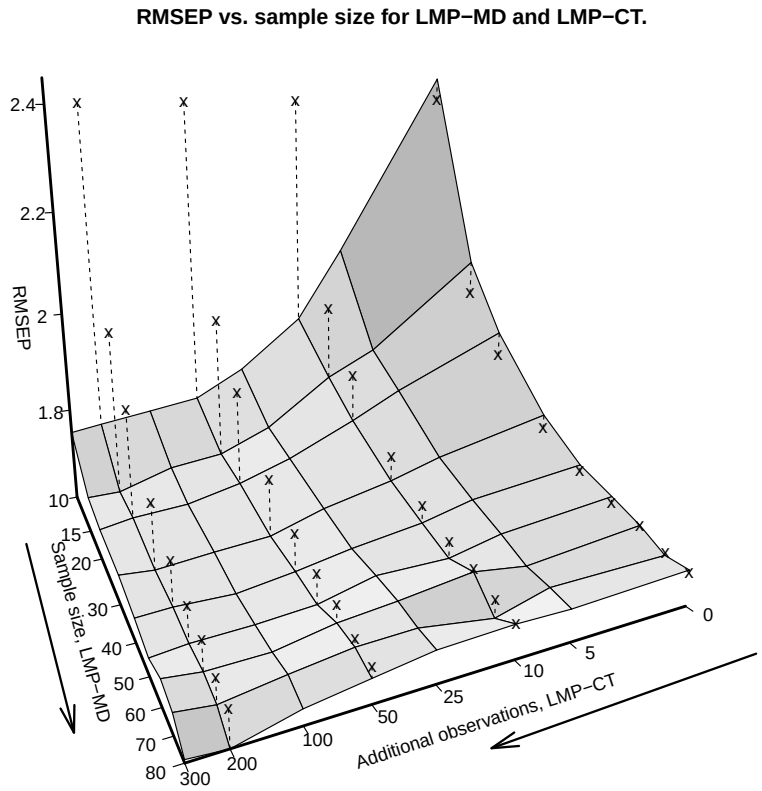


Figure 2: *REMSEP* for LMP–MD based on samples with varying number of observations for LMP–MD and additional observations of LMP–CT. Each sample size combination for LMP–MD and LMP–CT represent a "line–crossing" in the plane, and is represented with the average *REMSEP* value for 500 (1000 for LMP–MD size at 80) different random samples of LMP–MD and LMP–CT using the empirical Bayes model. Points represented by "x"–s show corresponding *REMSEP*s using OLS. The dotted lines represent the difference (i.e. the gain) of using the empirical Bayes method instead of OLS for each combination of sample–sizes for LMP–MD and LMP–CT.

specially design to reduce prediction precision. Thus the parameter–estimates, both for $\boldsymbol{\beta}$ and for $\boldsymbol{\Sigma}$ are (very mildly) biased.

The conditional variance, i.e. conditional on known regression parameter and predictor variables, of LMP–MD (σ_{22}) was smaller than the conditional variance of LMP–CT (σ_{11}) as a 95% Credible Interval for the fraction σ_{22}/σ_{11} was given by (0.57, 0.79), with a posterior median at 0.67. This corresponds to higher observed RMSEP–values for LMP–CT, typically around 1.9 for the best models.

3.2 Sufficient sample size

Figure 2 shows the relationship between RMSEP and sample sizes for LMP–MD and LMP–CT, using the two methods OLS and empirical Bayes. The figure shows that when there

were no additional observations for LMP–CT, OLS generated regression parameter estimates just as good, or even better than the empirical Bayes method. However, when additional observations of LMP–CT were present the empirical Bayes method generated regression parameter estimates yielding increased prediction precision compared with OLS–estimates.

4 Discussion

As the data structure with more observations for LMP–CT than LMP–MD is likely to be a challenge in numerous countries inside EU it might be advantageous to include methods for analysing such data in Causeur et al. (2003). A prerequisite for such inclusion is that there exist suitable software, like R–packages. Unfortunately a R–package for the empirical Bayes method applied in this study is not yet available. For 2SLS and random effects models some such software exists (Bates et al., 2015; Henningsen and Hamann, 2007; Pinheiro et al., 2015). Diggle et al. (e.g. 2002, chap 4.5) outlines a method for analysing the model in (1) by frequentist principles.

The variation between regression parameter estimates for all methods, including OLS using LMP–MD as a single response, was close to negligible. The favored model, i.e. the model using HGP7–*variables* and *Gender* as predictor variables stood out as the better model independently of method used for model selection (model evidense or RMSEP). The model omitting *Gender* as predictor variable performed almost as well as the favoured model, and has the advantage of simpler data–sampling since gender doesn't have to be registered.

The signs of the regression parameter estimates are in line with prior expectations. A thick *Fat* layer decreases predicted LMP, and oppositely a thick *Meat* layer increases expected LMP. The inner layer of fat (*Totif*) is the least important predictor variable from the HGP7. Its effect in the regression equation is likely to be severely affected by the other more dominant predictor variables. Finally the results show that conditional on similar measurements for HGP7–*variables*, females were expected to have a higher LMP than castrates.

When *Year* was included in the analysis the motive was not increased prediction precision as we have no observations for years 2014, 2015 etc. in the training data. Thus it is not possible to include *Year* as a predictor in the future. The motive for including *Year* in the analysis was to see if any effect was present, effects that are not a result of the *Year* itself. Such effects might have numerous explanations like differences in the manual– or CT–dissection (different butcher teams), development of the body composition in pigs

etc. As the effect of *Year* in the present study is close to negligible it indicates that the validity of the prediction equation is applicable over time. This interpretation is subject to considerable uncertainty as only two different years, 2008 and 2013, is involved in the study.

Omitting *Gender* as a predictor variable is in line with the principles used in the Netherlands, where the effect of gender on the prediction equation using HGP7 has been evaluated in detail by Engel et al. (2006, 2012). Engel et al. (2012) was a study aiming at finding robust methods for handling different proportions of females, castrates and males in the pig population. They found a significant effect of *Gender*, but ended up using a prediction equations for HGP7 where *Fat* and *Meat* were the two only predictor variables, data for the inner fat layer (*Totif*) was not used in their analysis. Engel et al. (2006, 2012) and Font i Furnols and Gispert (2009) reported RMSEP at 2.24, 2.10 and 1.8 % units respectively, a little larger than the RMSEP at 1.63, reported for Norway in the present study. In Font i Furnols and Gispert (2009) Fat–O–Meater, an optical probe similar to HGP7, was used.

The results from this study show a significant effect of *Gender* as predictor variable for LMP–MD. Due to considerable extra costs if *Gender* has to be sampled for every carcass, models omitting *Gender* might be preferable. Since the bulk of pig farmers deliver a close to equal proportion of females and castrates to slaughterhouses, omitting *Gender* as predictor variable will have minor effect on the total cash settlement between farmer and slaughterhouse. However, the omission will lead to a bias where LMP, in average, will be underestimated for females and overestimated for castrates.

The effect of *Gender* indicate differences between females and castrates regarding the meat and fat distribution in the carcass. If these differences affects the profitability of the possible processing methods, it might be profitable for slaughterhouses to register *Gender*. Further analysis of such effects falls outside the scope of this study.

Since the distribution of fat and meat in the carcasses to some extent depends on the gender, it is very likely that such dependencies might also occur between breeds, even if the effect was ignorable for the three breeds evaluated in the present study.

An eventual effect of *Breed*, unlike *Gender*, would introduce a bias providing a systematic effect, positive or negative, for different farmers and cooperatives using different breeds. Since LMP differs between breeds, see Appendix A.1, there is a consistent demand from different cooperatives and farmers that the effect of *Breed* is to be tested and accounted for in the prediction equation. Consequently the ignorable effect of *Breed* demonstrated for the present study does not rule out testing for this effect in forthcoming updates of the prediction equation for LMP.

Furthermore, there seems to be very limited gain in including interaction–terms. Such inclusions makes the models more complex, and thereby increases the possibility for substantially biased prediction of carcasses having anomalous on–line measurement values. The non–usefulness of second–order interaction terms strongly suggests that including higher order interaction terms or quadratic terms would not be beneficial.

The estimated values for the different elements of $\boldsymbol{\beta}$ were very similar for the estimates regarding LMP–MD and LMP–CT, with exception for the intercept–term, which was larger for LMP–CT. The natural interpretation is that this difference reflects the difference of 6.3 percent units between average observed LMP–CT and LMP–MD.

Statistically the expected value for RMSEP is the square root of the sum of *squared bias* and *error variance* for the predicted values. For any given model the only way to reduce the expected value for RMSEP is to reduce *squared bias* as the (model specific) *error variance* is assumed fixed.

In the present case, with 86 observations of LMP–MD, and a fairly simple model including only four predictor variables, the main part of prediction errors are due to the modelled error variance, and not a result of biased regression parameters. Consequently the empirical Bayes method did not substantially improve prediction precision for LMP–MD in terms of RMSEP.

The usefulness of applying the empirical Bayes method, or other methods utilizing the information from a surrogate variable, i.e. LMP–CT, depends heavily on the sample sizes of observed LMP–MD and LMP–CT, and the covariance–matrix for the error terms ($\boldsymbol{\Sigma}$). The relationship between sample sizes of manually dissected carcasses, CT–scanned carcasses and RMSEP shown in Figure 2 might be used to optimize sample sizes if the cost of sampling and the gain of reduced expected prediction error is known. The empirical Bayes method might be applied to heavily reduced sets of data without appreciable loss of prediction precision.

In situations where several breeds are present methods utilizing CT–data will be especially useful. Then the training set might be composed of a limited number of manual dissected carcasses, and a larger number of CT–scanned carcasses containing a sufficient number of carcasses from all breeds.

Prior to looking at the results we had expected higher precision for predicted LMP–CT values compared with LMP–MD values, due to an assumption that CT is a more precise method for predicting LMP than manual dissection. Evaluation of the manual dissection method shows a generally high accuracy and reliability for the estimated LMP–MD (Nissen et al., 2006), but also revealed some problematic issues, for instance a significant effect of butcher for the estimated LMP–MD.

The results in this study show the opposite, that on–line measurements using HGP7 tend to predict LMP–MD more precisely than LMP–CT. This is seen by evaluating the variances for the errors, and by comparing RMSEPs for LMP–MD and LMP–CT. The pattern is evident even if the sample is restricted to the carcasses where both LMP–MD and LMP–CT are observed. Thus, the non random sampling of carcasses for manual dissection does not explain the observation. However, the difference might, at least partly, be explained by the fact that the manual dissection method is a partial dissction method (Walstra and Merkus, 1995), whereas CT performs a total dissection. Thus the measurements obtained by the HGP7 on the back of the carcass might be better correlated with the partial manual dissection than the full CT–dissection.

The estimate for $\sigma_{22} = \sigma_{MD}$, and the estimated RMSEP values indicates that the EU specification of RMSEP at maximum 2,5 is easily fulfilled using HGP7 on Norwegian carcasses.

5 Conclusion

A model using four predictor variables; *Fat*, *Meat*, *Totif* (from the optical probe – HGP7) and *Gender*, is simple and provides a high prediction precision well inside EU standards. The three variables *Year*, *Breed* and *Weight* seems to be of none or minor importance for LMP–MD prediction when combined with the HGP7–variables and *Gender* for the data set used in the present study. This result should not be generalized without reservation. Inclusion of second–order interaction terms does not improve prediction precision substantially.

For analytical simplicity OLS–regression using input from the optical probe and gender is the better method and model. However, for practical simplicity gender might be omitted without severe loss of prediction precision.

The drawback of OLS–regression is its inability to utilize extra information from CT–scanned carcasses. In order to limit the needed sample size of manually dissected carcasses the empirical Bayes regression might be applied to a training sample with a low number of manually dissected carcasses, but a sufficiently large number of CT–scanned carcasses, yielding no or minor loss of prediction precision.

6 Funding

This study is part of the research project PigComp. PigComp is partly funded from the Research Council of Norway via the program "Innovation projects for the industrial sector" (projectnumber 225294). Other partners and contributors to the project are Animalia,

Nortura SA, Norsvin SA, Furuseth AS, the Norwegian Meat and Poultry Association (KLF) and the Norwegian University of Life Sciences.

7 Conflict of interest

All authors declare no conflict of interest.

8 Acknowledgments

Training schools and seminars arranged by the EU–founded COST Action FAIM has provided valuable knowledge for the present study. Trygve Almøy and Frøydis Bjerke has contributed to this article by useful comments to the manuscript. Finally thanks are due to all personnel at slaughterhouses, Animalia and DMRI as their patience and effort made the data sampling possible.

References

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting Linear Mixed–Effects Models Using lme4. *Journal of Statistical Software*, 67(1):1–48.

Carlin, B. P. and Louis, T. A. (2008). *Bayesian Methods for Data Analysis*. CRC Press.

Causeur, D., Daumas, G., Dhorne, T., Engel, B., Font i Furnols, M., and S, H. (2003). Statistical handbook for assessing pig classification methods: Recommendations from the "EUPIGCLASS" project group.

Commission of the European Communities (2006). Commission Regulation (EC) No 1197/2006 of 7 august 2006 amending Regulation (EEC) No 2967/85 laying down detailed rules for the application of the Community scale for grading pig carcases. *Official Journal of the European Union*, L 217(49):6–7.

Commission of the European Communities (2008). Commission Regulation (EC) No 1249/2008 of 10 December 2008 laying down detailed rules on the implementation of the Community scales for the classification of beef, pig and sheep carcases and the reporting of prices thereof. *Official Journal of the European Communities*, L 337:3–30.

Daumas, G. and Monziols, M. (2011). Comparison between computed tomography and dissection for calibrating pig classification methods. In *57th International Congress of*

Meat Science and Technology (ICoMST 2011), volume 1, pages 296–299. Ghent-Belgium, 7–12 August 2011.

Diggle, P. J., Liang, K.-Y., and Zeger, S. L. (2002). *Analysis of Longitudinal Data*. Oxford University Press, Oxford,UK.

Engel, B., Lambooij, E., Buist, W., Reimert, H., and Mateman, G. (2006). Prediction of the percentage lean of pig carcasses with a small or a large number of instrumental carcass measurements – an illustration with HGP and Vision. *Animal Science*, 82(6):919–928.

Engel, B., Lambooij, E., Buist, W., and Vereijken, P. (2012). Lean meat prediction with HGP, CGM and CSB-Image-Meater, with prediction accuracy evaluated for different proportions of gilts, boars and castrated boars in the pig population. *Meat science*, 90(2):338–344.

Font i Furnols, M. and Gispert, M. (2009). Comparison of different devices for predicting the lean meat percentage of pig carcasses. *Meat science*, 83(3):443–446.

Gangsei, L. E., Almøy, T., and Sæbø, S. (2016). Linear regression with bivariate response variable containing missing data. An empirical Bayes strategy to increase prediction precision. Submitted manuscript to Communications in Statistics – Simulation and Computation.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996). Introducing Markov Chain Monte Carlo. In Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., editors, *Markov Chain Monte Carlo in Practice*, chapter 1, pages 1–19. London: Chapman and Hall.

Henningsen, A. and Hamann, J. D. (2007). systemfit: A Package for Estimating Systems of Simultaneous Equations in R. *Journal of Statistical Software*, 23(4):1–40.

Judas, M., Höereth, R., and Branscheid, W. (2007). Computed tomography as a method to analyse the tissue composition of pig carcases. *Fleischwirtschaft International*, 1:56–59.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *J. Amer. Statist. Assoc.*, 90(430):773–795.

McLeod, A. and Xu, C. (2011). *bestglm: Best Subset GLM*. R package version 0.33.

Nissen, P. M., Busk, H., Oksama, M., Seynaeve, M., Gispert, M., Walstra, P., Hansson, I., and Olsen, E. (2006). The estimated accuracy of the EU reference dissection method for pig carcass classification. *Meat science*, 73(1):22–28.

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D., and R Core Team (2015). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-121.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The annals of statistics*, 6(2):461–464.

Vester-Christensen, M., Erbou, S. G., Hansen, M. F., Olsen, E. V., Christensen, L. B., Hviid, M., Ersbøll, B. K., and Larsen, R. (2009). Virtual dissection of pig carcasses. *Meat science*, 81(4):699–704.

Walstra, P. and Merkus, G. (1995). *Procedure for the assessment of lean meat percentage as a consequence of the new EU reference dissection method in pig carcass classification*. DLO Research Institute of Animal Science and Health (ID-DLO), Zeist, The Netherlands.

Wooldridge, J. M. (2012). *Introductory Econometrics: A Modern Approach*. Cengage Learning.

A Appendix

A.1 Summary statistics for response and predictor variables

Table A1: Mean and standard deviation, in parenthesis, for LMP–MD (%), LMP–CT (%) and the 4 continuous predictor variables (*Fat* (mm), *Meat* (mm), *Totif* (mm) and *Weight* (kg)). Values distributed over the two levels for *Year*, i.e. 2008 and 2013, and the three levels for *Breed*, i.e. Hampshire, Norhybrid and Duroc. The numbers n_{CT} and n_{MD} shows sample sizes for CT–scanned and manual dissected carcasses respectively.

2008	**Hamp.**	**Norhyb.**	**Duroc**	**All**
n_{CT}	37	75	117	229
n_{MD}	10	20	36	66
LMP–CT	66.6 (3.1)	66.5 (4.1)	66.4 (3.3)	66.4 (3.5)
LMP–MD	60.5 (3.6)	60.6 (4.1)	60.1 (2.4)	60.3 (3.1)
Fat	11.9 (2.1)	12.5 (3)	12.4 (2.3)	12.3 (2.5)
Meat	57.9 (6.8)	54.9 (5.5)	56.1 (6)	56 (6)
Totif	12.7 (4.9)	12.6 (3)	10.6 (4.3)	11.6 (4.2)
Weight	81.1 (6)	82.1 (6.8)	78.4 (5.8)	80.1 (6.4)
2013	**Hamp.**	**Norhyb.**	**Duroc**	**All**
n_{CT}	32	78	126	236
n_{MD}	10	0	10	20
LMP–CT	66.6 (2.6)	67.5 (4)	66.3 (3.7)	66.7 (3.7)
LMP–MD	60.6 (2.8)	–	59.8 (3.5)	60.2 (3.1)
Fat	11.9 (2.2)	11.3 (2.8)	12.2 (2.7)	11.8 (2.7)
Meat	59.6 (6.2)	55.7 (6.1)	53.3 (7.1)	55 (7)
Totif	10.8 (3.7)	13 (4.2)	10.6 (3.9)	11.4 (4.1)
Weight	80.8 (6.6)	79.2 (8.6)	76.1 (7.6)	77.7 (8)
Both years	**Hamp.**	**Norhyb.**	**Duroc**	**All**
n_{CT}	69	153	243	465
n_{MD}	20	20	46	86
LMP–CT	66.6 (2.9)	67 (4.1)	66.3 (3.5)	66.6 (3.6)
LMP–MD	60.5 (3.1)	60.6 (4.1)	60.1 (2.6)	60.3 (3.1)
Fat	11.9 (2.2)	11.9 (3)	12.3 (2.5)	12.1 (2.7)
Meat	58.7 (6.5)	55.3 (5.8)	54.7 (6.7)	55.5 (6.5)
Totif	11.8 (4.5)	12.8 (3.7)	10.6 (4.1)	11.5 (4.2)
Weight	80.9 (6.2)	80.6 (7.9)	77.2 (6.9)	78.9 (7.3)

Paper IV

Computers and Electronics in Agriculture, Vol.121, pp.191–194

doi:10.1016/j.compag.2015.12.002

Automatic Segmentation of Computed Tomography (CT) images of domestic pig skeleton using a 3D expansion of Dijkstra's algorithm.

Lars Erik Gangsei[a] & Jørgen Kongsro[b]

Abstract: A 3D expansion of Dijkstra's algorithm used for automatic segmentation and identification of the bones in CT images of live pigs was developed and validated. The major bones in the skeletons of 208 out of 485 live pigs (43%) were segmented and identified from the images without major errors. The segmentation and identification is executed through 8 main operations: (1) identify the full bone structure by a threshold of Hounsfield units, (2) identify forelimbs by voxel connectivity and set landmarks, (3 - 8) segment out and identify the individual bones in different main parts of the bone structure by the 3D expansion of Dijkstra's algorithm. The algorithms described will constitute an important basis for further work applying CT in pig breeding and management.

Keywords: Computer Tomography; pig; skeleton; segmentation

a: Animalia, P.O. Box 396 - Økern, N-0513 Oslo, Norway

b: Norsvin SA, P.O. Box 504, N-2304 Hamar, Norway

1 Introduction

Computer Tomography (CT) is a highly suitable method for identifying bonestructures inside living and dead organisms containing a calcified skeletal structure, typically mammals and fish. The full skeleton structure might easily be roughly segmented out from the rest of body tissue by adding a threshold value, typically around 200 Hounsfield units (HU) (Fiebich et al., 1999), and by classifying all voxels with larger values as bone.

The ability to segment the carcass into identifiable bones has several purposes. Our ultimate goal is to construct a complete labeled body atlas for pigs. Segmented and identified bones from a plethora of pigs will be indispensable in the process of constructing the atlas, as the skeleton will constitute the framework of the atlas. The atlas made for mice by Dogdas et al. (2007), is an illustrative example for what we want to achieve. The bone segmentation might also be used for other purposes, for instance to diagnose diseases or undesirable qualities in the skeleton.

By a method known as atlas-based segmentation (Baiker et al., 2010; Cuadra et al., 2015), the different parts of the pig, i.e. the CT scan, might be identified as cuts, muscles, organs etc. The highest priority is to be able to identify the main commercial cuts in live animals.

The aim was to describe an approach for automatic segmentation and identification of the larger bones from CT images of live domestic pigs. To our knowledge, this is the first automatic algorithm for segmenting pig skeletons in a volume generated from CT images.

2 Material and Methods

2.1 Animals and the ct scan

The material consists of in vivo CT scans from 485 boars at the Norsvin Delta test station for purebred boars (Norsvin SA). The pigs were purebred Norsvin Landrace and Duroc boars. The live weight is as close to 120 kg as practically possible. This is due to 120 kg representing the end point of the testing period for terminal boars, and is regarded as the optimal carcass weight (70-80 kg) in Norway. The samples were selected randomly from the annual breeding stock of 3500 boars tested in the Norsvin breeding program. The CT scanner was a GE Healthcare LightSpeed 32 VCT, and the settings used were 120kV, slice thickness 1.25 mm and dynamic mA (400-500 mA) adjusting for object thickness.

Prior to CT scanning, the boars were sedated using Azaperone, 8mg/kg i.m. (Stresnil Vet R , Janssen-Cilag Ltd, Buckinghamshire, UK). Boars were scanned, 45 min after injection,

doi:10.1016/j.compag.2015.12.002

as the sedation was given to help facilitate the scanning procedure and improve image quality. All animals were cared for according to the laws and regulations for keeping pigs in Norway (Aasmundstad et al., 2013). After sedation, the pigs were transported to the CT scanner using a crib made of fiberglass.

The software MATLAB (MATLAB, 2014), including the Image Processing Toolbox, was used for the segmentation. The code for landmark identification and the 3D expansion of Dijkstra's algorithm was written from scratch. MATLAB code for a function conducting the 3D expansion of Dijkstra's algorithm will be provided by authors upon request.

2.2 General methods for segmentation

We applied three main principles for identifying and segmenting the individual bones; (i) segmentation by connectivity, (ii) identification of points and lines and (iii) a 3D expansion of Dijkstra's algorithm.

Segmentation by connectivity was primarily involved at the start of the process. A binary 3D image was constructed by applying a threshold value for bone (HU $>$ 180). The different connected objects in this 3D binary image were labeled by the MATLAB function "bwconncomp", which takes a binary image as input and returns an image where each connected object is labelled with a specific value. The connected objects inside the 3D binary image were identified by ranking their volumes or mass center points. The 4 largest connected objects from a randomly selected pig are shown in Fig. 1.

To identify landmarks, 2D projections of the skeleton were used extensively. The principles are explained through an example illustrated in Fig. 2. The steps were as follows:

1. A 2D image containing the sums of bone-voxels perpendicular to the sagittal plane (in this example) was constructed, i.e. high intensity areas represent areas with thick bones. One point, approximately corresponding to *Trochanter major*, was already known from similar techniques, as illustrated in Fig. 2a.

2. Based on the known point, a new region of interest (ROI) was defined based on euclidian distances. Points closer than 50 mm, a figure set a priori based on experience, from the known point were "masked out", (Fig. 2b).

3. A new point was set at the point having maximal intensity inside the new ROI (Fig. 2c).

4. Steps similar to steps 2-3 were redone to find the rest of the points of interest resulting in a set of four "known fixed points" as illustrated in Fig. 2d.

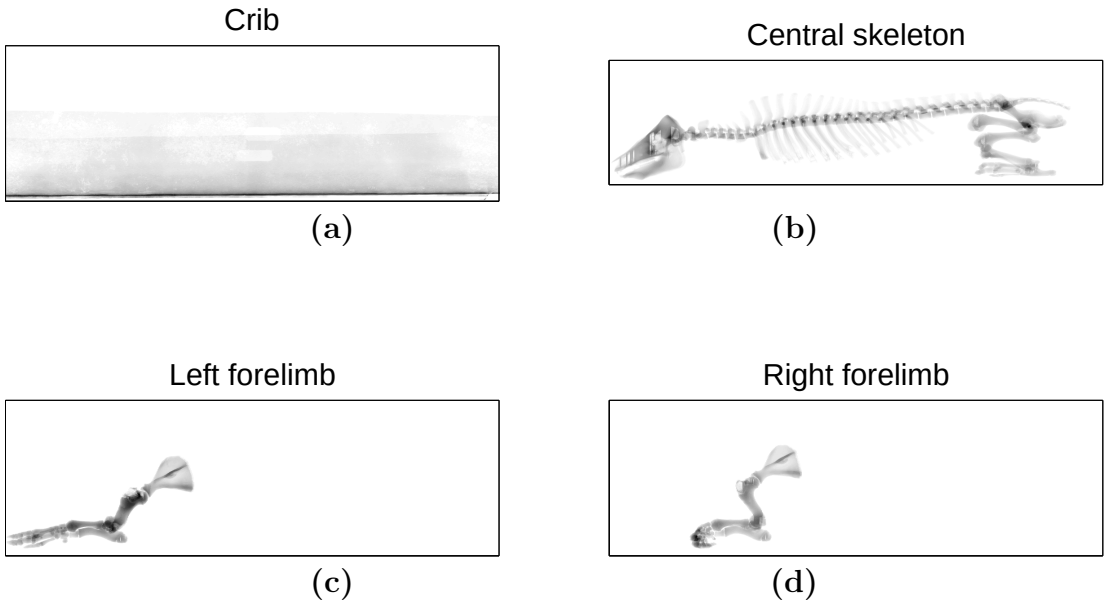


Figure 1: An example of segmentation thorough connectivity. The 4 panels represents the 4 largest connected objects, i.e. crib (a), "central skeleton" (b), left forelimb (c), and right forelimb (d). The input is a binary image produced by setting a threshold at 180 HU for a CT image of a random pig.

5. A variant of this method is to define the ROI as a line, for instance the upper part of the spine, by using the MATLAB function "bwboundaries". The points on top of each vertebra and points between vertebras are set successively at local maximum and minimum sagittal values for this line.

Most of the bones had to be segmented and treated one by one. Dijkstra's algorithm (Dijkstra, 1959), a basic method of image analysis, is well known for identifying the cheapest path between two nodes, and is thereby a suitable method for separating different objects in a 2D image. The method we applied might be viewed as an expansion of Dijkstra's algorithm in order to separate a 3D, not 2D, object.

The first step was to identify a ROI (3D array) that encapsulates the surface in which the two bones were connected. These ROIs were automatically constructed based on the identified landmarks and prior information on bone dimensions and the spatial orientation of bones. As an example, the ROI illustrated by the box in Fig. 3 was constructed via a landmark in the top of the overarm, i.e. the lower right corner of the box. The horizontal distance for the ROI was set by prior knowledge. The two remaining dimensions of the ROI were easily set by the extreme values for bone voxels within the horizontal limitation.

In the next paragraphs, we will view the 3D binary input array (bone/ not bone) inside

DOI:10.1016/j.compag.2015.12.002

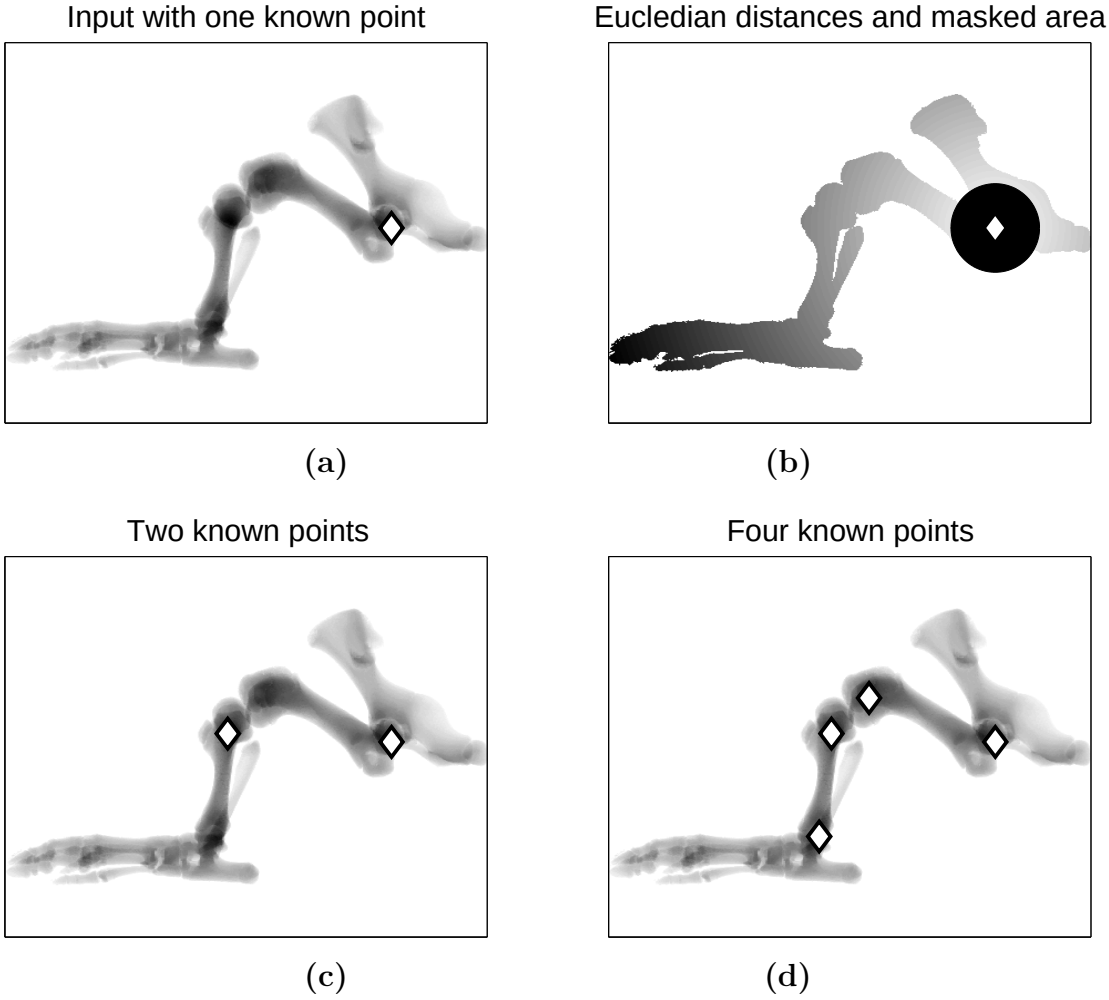


Figure 2: Principles for identifying points in hindlimbs. All panels show a right hindlimb observed perpendicular to the sagittal plane. The intensities in (a), (c) and (d) represent the sum of bone voxels over the full 3D array. The intensities in (b) represent the euclidian distances to the "known point" ($\approx$*Trochanter major*).

the ROI as a stack of 2D layers. For each layer, a virtual cut is constructed in the "cutting direction". The sum of these cuts constitutes the segmentation surface. The directions of the layers and the cutting direction varied with the main direction of the segmentation surface, and were set based on prior knowledge. For instance, vertebras were split approximately parallel to the transverse plane, pelvis was segmented from spine approximately parallel to the sagittal plane etc. When the segmentation surface is approximately parallel to the transverse plane as shown in Fig. 3, the layer direction is from top to bottom and cutting direction is from the right to left side of the pig. The two bones to be segmented had to penetrate the two opposite sides of the input array orientated perpendicular to both cutting- and layer-direction.

The first step of Dijkstra's algorithm (2D) is to construct a cost matrix, i.e. a matrix defining the minimum cumulative cost of including any voxel to the virtual cut. The cost

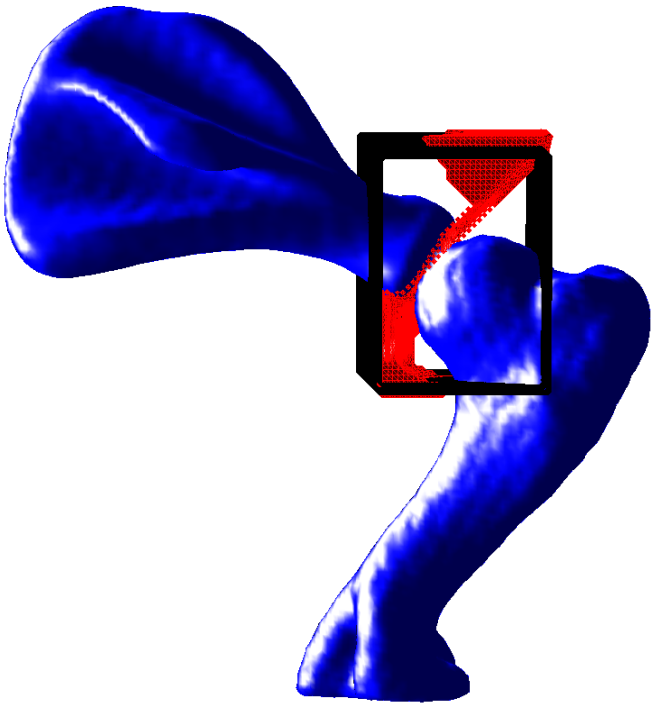


Figure 3: 3D representations of shoulderblade and overarm. The "black box" represents the ROI used in the segmentation. The red surface show the surface where the bones are separated.

matrix for all layers was constructed by the standard Dijkstra's algorithm. For the first layer, the input to the algorithm was the first layer of the binary input array. For the next layers the input consisted of the sum of the cost matrix from the previous layer and the corresponding layer of the input array. The final 3D cost array was the array of stacked 2D cost matrices.

The next step in Dijkstra's algorithm is to construct the cheapest path, i.e. the segmentation surface. For all layers, these paths were constructed via the standard 2D Dijkstra's algorithm based on the corresponding layers from the 3D cost array. The algorithm started with the last layer. After the path in one layer was identified, a region of "possible paths", i.e points connected to the identified path, was identified for the next layer. All points outside this possible path in the 3D cost array were set to infinity. This step was repeated successively for all layers. This method ensured that the resulting separating surface minimized the total cost for separating the two bones, and that the surface was connected in both the "cutting direction" and along the layer direction as illustrated in Fig. 3.

doi:10.1016/j.compag.2015.12.002

2.3 Work flow

The work flow we applied is described through the following enumerated list. The steps of the graphical abstract (online version only) correspond to this numbering.

1. Upload CT image and segment out the skeleton by threshold (180HU).
2. Split the skeleton into three major parts, two forelimbs and "central skeleton" by connectivity.
3. Segment the two forelimbs into three parts: shoulderblade (*Scapula*), overarm (*Humerus*) and "the rest".
4. Segment the two hindlimbs from the rest of the "central skeleton".
5. Split the two hindlimbs into 4 smaller parts: pelvic girdle (*Illium*, *Pubis* and *Ischium*), leg bone (*Femur*), rear leg bones (*Fibula* and *Tibia*), and foot, (Tarsal bones and Metatarsal bones).
6. Segment the rest of the "central skeleton" into 7 main parts: head (including jaw), breast bone (*Sternum*), neck vertebras (*Vertebrae cervicis*), thoracic vertebras (*Vertebrae thoracis*), ribs (*costae*), lower back vertebras (*Vertebrae lumbalis*) including sacrum (*Os sacrum*) and tail (*Vertebrae coccygis*).
7. Segment and identify the individual ribs.
8. Segment the neck vertebras, thoracic vertebras, and lower back vertebras into single identified vertebras.

3 Results and discussion

We were not able to find a stable solution for segmenting out the foreshank bones (*Radius* and *Ulna*) and foot (carpal bones and metacarpal bones), in the forelimbs. We did not prioritize finding an algorithm for these parts as the prime meat cuts surrounding these bones are of subordinate interest.

We applied the algorithm to 485 pigs. The results of the segmentation were evaluated by visual inspection. The main reason for segmentation and identification failure was misplaced landmarks. In order to reduce the fraction of skeletons that are unsatisfactorily segmented, some semi-automatic methods, for instance visual inspection and correction of landmarks, might be necessary to implement.

If the segmentation procedure failed for a pig at a given step in the work flow, the remaining steps were excluded as such failures heavily complicate or even preclude the consecutive steps of the segmentation process. The failure rates for different steps are summed up in the following results.

- 2 out of 485 (< 1%) pigs failed in the first step segmenting out the main part of the skeleton.
- 17 of (the remaining) 483 (3,5%) pigs failed in identifying points in the hindlimbs.
- 65 of 466 (14%) failed in segmenting the vertebra in its main parts, mainly identifying the points where last rib meets the vertebra.
- 25 of 401 (6%) failed in the segmentation of bones in the hindlimbs.
- 9 of 376 (2%) failed when identifying points in forelimbs.
- 67 of 367 (18%) failed in segmenting and identifying the ribs.
- 37 of 300 (12%) failed in segmenting the lower back vertebras to individual vertebras.
- 43 of 263 (16%) failed in identifying the last lower back vertebra.
- 12 of 220 (5%) failed in the segmentation of the forelimb.
- Totally 208 of 485 (43%) pig skeletons were segmented and identified without errors.

There is a wide scope of opportunities to improve the algorithm with respect to stability and speed, e.g. by GPU computation, within the principles outlined in this application. The "expanded" Dijkstra's algorithm might be further developed. For instance by using continuous (intensities, gradients etc.), not binary variables, in the input, or by expanding the possibilities for how the resulting surface is allowed to be connected.

4 Conclusion

The approach described in this application might be further developed. However, the main principles will constitute an important basis for further work analysing CT scans of pigs.

5 Acknowledgements

We want to thank Ole Alvseike for proof reading and for making valuable suggestions to this article.

DOI:10.1016/j.compag.2015.12.002

6 References

References

Aasmundstad, T., Kongsro, J., Wetten, M., Dolvik, N., and Vangen, O. (2013). Osteochondrosis in pigs diagnosed with computed tomography: heritabilities and genetic correlations to weight gain in specific age intervals. *Animal: An International Journal of Animal Bioscience*, 7(10):1576–1582.

Baiker, M., Milles, J., Dijkstra, J., Henning, T. D., Weber, A. W., Que, I., Kaijzel, E. L., Löwik, C. W., Reiber, J. H., and Lelieveldt, B. P. (2010). Atlas-based whole-body segmentation of mice from low-contrast Micro-CT data. *Medical Image Analysis*, 14(6):723–737.

Cuadra, M. B., Duay, V., and Thiran, J.-P. (2015). Atlas-based Segmentation. In *Handbook of Biomedical Imaging*, pages 221–244. Springer.

Dijkstra, E. W. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271.

Dogdas, B., Stout, D., Chatziioannou, A. F., and Leahy, R. M. (2007). Digimouse: a 3D whole body mouse atlas from CT and cryosection data. *Physics in Medicine and Biology*, 52(3):577.

Fiebich, M., Straus, C. M., Sehgal, V., Renger, B. C., Doi, K., and Hoffmann, K. R. (1999). Automatic bone segmentation technique for ct angiographic studies. *Journal of computer assisted tomography*, 23(1):155–161.

MATLAB (2014). *version 8.3.0.532 (R2014a)*. The MathWorks Inc., Natick, Massachusetts.

Computers and Electronics in Agriculture, Vol.121, pp.191–194

doi:10.1016/j.compag.2015.12.002

Paper V

Building an *in vivo* anatomical atlas to close the phenomic gap in animal breeding.

Lars Erik Gangsei[a,c], Jørgen Kongsro[b],Kristin Olstad[d], Eli Grindflek[b] & Solve Sæbø[c]

Abstract: Currently, a growing gap is observed between the enormous amount of genomic information generated from genotyping and sequencing and the scale and quality of phenotypes in animal breeding. In order to fill this gap, new technologies and automated large-scale measurements are needed. Body composition is an important trait in animal breeding related to growth, feed efficiency, health, meat quality and market value of farmed animals. *In vivo* anatomical atlases from CT will aid large-scale and high-throughput phenotyping in order to reduce some of the gap between genotyping and phenotyping in animal breeding. We demonstrated that atlas segmentation was able to predict major parts and organs of the pig with a numerical test applied to the primal commercial cuts.

a: Animalia, P.O. Box 396 - Økern, N-0513 Oslo, Norway

b: Norsvin SA, P.O. Box 504, N-2304 Hamar, Norway

c: Norwegian University of Life Sciences (NMBU), Department of Chemistry, Biotechnology and Food Science, P.O. Box 5003, N-1432 Ås, Norway

d: Norwegian University of Life Sciences, Department of Companion Animal Clinical Sciences, Equine Section, N-0454 Oslo, Norway

1 Introduction

Recent advances in genome sequencing technology has led to high-throughput and high–density information in humans, animals and plants (Houle et al., 2010). Variation in phenotypes is produced through a web of interactions between genotype and environment, and there is a need for detailed phenotypic data to characterize the phenomes. Measuring body composition in farmed animal breeding is important in order to improve growth and feed efficiency, health, meat quality and market value of carcasses. Body composition has traditionally been assessed by a number of different means, ranging from subjective scoring or simple point measurements of subcutaneous fat to physical dissection of carcasses or *in vivo* volume scans using Computed Tomography (CT) or Magnetic Resonance Imaging (MRI).

For pigs, the use of CT makes it possible to obtain accurate *in vivo* measurements of body composition (Gjerlaug-Enger et al., 2012). Genetic selection on body composition traits in pigs was previously done by physical dissection of full-sibs and half-sibs of the selection candidates, which give much less accurate breeding value estimations compared with measuring body composition on the selection candidates themselves *in vivo*. Today, the pig breeding company Topigs Norsvin uses CT to measure body composition and monitor orthopedic disorders on 3.500 nucleus boars annually as an integrated part of their testing system. In this paper, we present an anatomical atlas from CT, which will help to close the phenomic gap in pig anatomy by giving access to high-throughput and high-dimensional anatomical phenotypes.

Obtaining *in vivo* body composition data from CT relies on segmentation of cross sectional slices. The segmentation strategies can be based on (1) intensities, applying adaptive thresholding of different tissues like adipose (fat), muscle and bone tissue, (2) shape or position using deformable models or active contours, and (3) labelled atlas (Commowick, 207). Methods are here ranked by complexity and demands of prior knowledge either from own data or literature. Automation of the segmentation methods would allow for detailed population studies of body composition. For atlas based segmentation, this paper show how an atlas can be constructed using a subset of animals from the population of pigs.

The atlas can serve as a framework for building large data sets of anatomical phenotypes, paving the way to detailed and high–density phenotypic information on pig anatomical traits. The number of additional variables in the breeding value estimation may be a limitation in terms of speed and complexity. The atlas phenotypes will be highly beneficial in terms of selection for animals with competitive advantages on muscle types, compared with the current selection in most breeding programs today, selecting for muscle- and fat

depth only. Creating atlases for primal cuts; "shoulder", "belly", "loin" and "ham", representing the market needs around the world would also make us able to sort our genetic material of pigs more efficiently in terms of different markets. Furthermore, by enhancing the anatomical traits by automatic segmentation, the accuracy of genetic selection for carcass traits will increase even further. The indirect effect of this is that more weight can be put in the breeding goal for hard-to-measure, low–heritable traits like maternal and disease-related traits, and in the end the whole breeding goal and genetic engine towards developing a more sustainable and accurate breeding program for farmed animals.

2 Methods

2.1 Approvement of the experiments

All animals were cared for according to laws, internationally recognized guidelines and regulations controlling experiments with live animals in Norway (Animal welfare Act 2009-06-19-97 (in Norwegian), 2009; Regulation for the keeping of pigs in Norway 2003-02-18-175 (in Norwegian), 2003); according to the rules given by Norwegian Animal Research Authority.

2.2 Data

The intensity atlas is in principle is the average of 386 nucleus boars, involving a total of approximately 3.4×10^{10} voxels (the 3D basic unit of the CT scans). The method was motivated by methods applied to micro CT scans of mice (Baiker et al., 2010; Li et al., 2008), where the skeletons were utilized as a framework for conducting the transformations.

The raw CT scans were volume representations of the individual pigs. The size of 3D data arrays (volumes) were approximately $512 \times 512 \times 1200$, where the third dimension, size, varied slightly with pig length. Each data point represented a voxel with size $0.9355mm \times 0.9355mm \times 1.25mm$. A CT intensity according to the Hounsfield (HU) scale was associated with each voxel.

2.3 Atlas

The atlas represents the average pig. The atlas volume size was $500 \times 500 \times 1600$, where each voxel represents a cube with a side length of 1 mm. We use the expressions "intensity atlas" and "labelled atlas", where the intensities aligned to each voxel might be interpreted

as HU–units. In the labelled version, every voxel is aligned to a specific label, i.e. organ, cut part etc.

Labelled and intensity volumes (3D) might be defined by a matrix representation, where the $N_y \times 3$ matrix $\mathbf{Y}$ and $N_x \times 3$ matrix $\mathbf{X}$, represent the atlas, and a random individual pig, respectively. N_y and N_x are the number of voxels in the respective images. Each row in $\mathbf{Y}$ and $\mathbf{X}$ defines the (Cartesian) coordinates for one voxel. The atlas was constructed through successive operations described in the next sections. Figures are used extensively to highlight important principles.

2.4 Skeleton atlas – image moments invariants

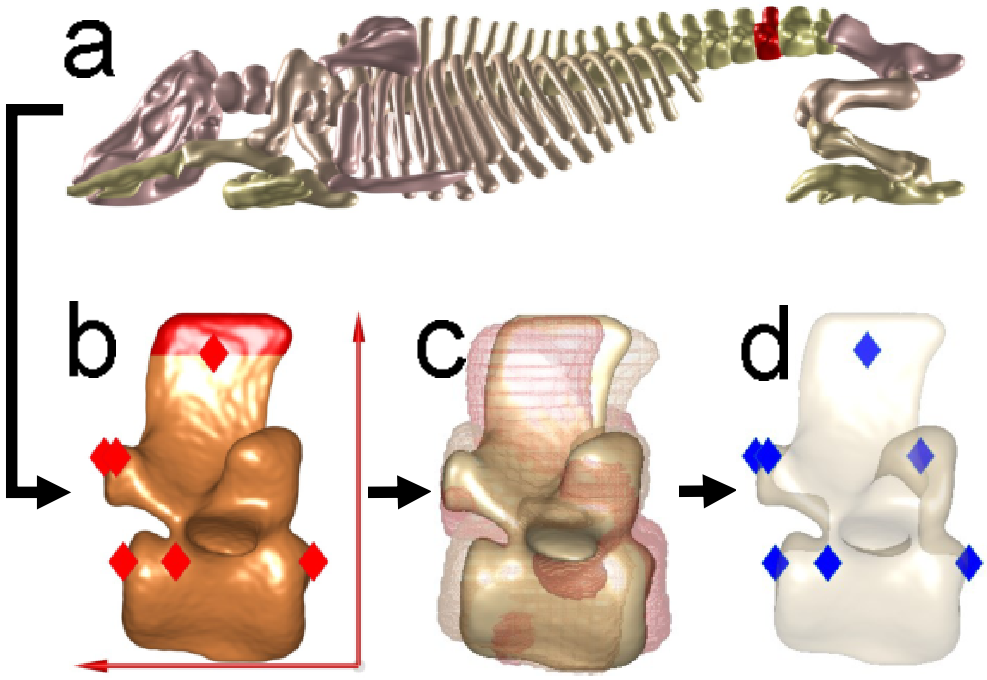


Figure 1: Construction of average bone by image moment invariants. (a) Segmented skeleton in a random pig; the vertebra illustrated in panels b–d is highlighted in red. (b) A vertebra with its orthonormal basis (arrows), landmarks, and the area where extra weight for orientation is added (red at top). (c) Construction of the average shape by rotating and scaling bones from all pigs to a common formwork. (d) Landmarks (blue) on the average vertebrae.

The first step was to identify the major bones in all pigs (Gangsei and Kongsro, 2016) (Fig. 1a). We calculated basic features for each bone, often referred to as image moments invariants (Hu, 1962): Center of mass (COM or $\bar{\mathbf{x}}$), the orthonormal basis of the bone ($\mathbf{R}$), volume ($v = n_\delta \times 0.9355^2 \times 1.25$, where n_δ is number of voxels) and length (l), that is, the Euclidian distance spanned by the bone along the first orthogonal basis vector. Left side bones were treated as right side bones by mirroring them over the sagittal plane before calculating the image moments invariants. The coordinates of each bone were represented

by the $n_\delta \times 3$ matrix $\mathbf{X}_\delta$. Furthermore, the diagonal weight matrix $\mathbf{W}$ assigned a specific weight to each voxel for the purpose of controlling the main directions of the orthonormal basis. The mathematical expressions for the COM and orthonormal basis were:

$$\bar{\mathbf{x}} = (1/n_\delta)\ \mathbf{X}_\delta^t \mathbf{1}_{n_\delta}, \quad \mathbf{R} = Eig\left\{ \left(\mathbf{X}_\delta - \mathbf{1}_{n_\delta}\bar{\mathbf{x}}^t\right)^t \mathbf{W}^2 \left(\mathbf{X}_\delta - \mathbf{1}_{n_\delta}\bar{\mathbf{x}}^t\right) \right\},$$

where the notation $Eig\{\mathbf{A}\}$ denotes the eigenvectors of the matrix $\mathbf{A}$ scaled to unit length.

The concept of the weighting of voxels is shown in Figure 1b, where the voxels in the red area, i.e. the voxels within a distance less than 1/10 of the total length (l) from the top, were given heavy weights (100). Thus, the first column in $\mathbf{R}$, i.e. the eigenvector having the largest corresponding eigenvalue, points approximately perpendicular to the coronal plane (upwards), the second eigenvector points approximately perpendicular to the transverse plane (forwards) and the third eigenvector points approximately perpendicular to the sagittal plane (to the left). For other bones, different parts were assigned additional weights, but the basic principle remains unchanged.

Based on the features of the individual bones we constructed atlas bones, i.e. templates for every bone in a pig (Fig. 1d). To every atlas bone, COM, volume, length, a common orthonormal basis and a shape, was applied. The COM ($\bar{\mathbf{x}}_T$), volume (v_T) and length (l_T) was just the average for all bones. For all bones in the spine and sternum, the COM value for the direction perpendicular to the sagittal plane (i.e. sideways), was set to 250 (mm). The common orthonormal basis, $\mathbf{R}_T$, was set to the individual orthonormal basis closest to the geometrically average ortonormal basis. Hence, by letting r_{ij} denote the element of the ith row and jth column of $\mathbf{R}$, and letting $\bar{r}_{ij}$ denote the average of the same element in all pigs, the $\mathbf{R}$ for which $\sum_{i=1}^{3}\sum_{j=1}^{3}(r_{ij} - \bar{r}_{ij})^2$ had the minimum value was chosen as the common orthonormal basis for the bone in question.

In order to construct the average shape, all bones were transformed to a 3D image, $\mathbf{B}$, of predefined size, $m_1 \times m_2 \times m_3$, (Fig 1c). The coordinates for the individual bones in these 3D images, denoted $\mathbf{Z}_\delta$, were given by rounded and scaled values of $(m_1/l)\ \mathbf{X}_\delta\mathbf{R}$. The scaling of $\mathbf{Z}_\delta$ was done by subtracting column means and adding column minimum values. Thus, every bone spanned the first dimension of $\mathbf{B}$ completely and was centred according to the two remaining dimensions. The final intensities of $\mathbf{B}$ equalled the sum of all bones transformed into it. The average shape was constructed by setting a threshold making sure that the volume of voxels in $\mathbf{B}$ having higher intensity than this threshold, was equal to the average volume of the bone (v_T).

2.5 Corresponding landmarks

The crucial steps of the method involved constructing corresponding landmarks between the volumes of the individual pigs (Fig. 2a–c). The initial step (Fig. 1d), was to set landmarks at approximately every 20mm along the main direction of the orthonormal basis of the average shaped bone. The landmarks were set either at the top, bottom, right and left side of the surface or in the centre of the bone (typically for ribs, hand and foot). In total approximately 1200 landmarks on the skeleton were identified (Fig.2a), varying with the number of vertebras and ribs in the individual pigs. The coordinates of the landmarks in the common orthogonal basis, $\mathbf{R}_T$, are denoted $\mathbf{Z}_l$, and the corresponding COM is denoted $\bar{\mathbf{z}}$.

These landmarks were transformed back to the basis of the individual pigs and the atlas by reversing the transformations based on image moments invariants. The common averages were used for the transformation to the atlas space resulting in a pattern symmetric over the sagittal plane (Fig. 2b). Individual image moments invariants were used for the individual pigs; consequently there was no symmetric pattern for these points (Fig. 2a). The mathematical expressions for the reverse transformations are given by:

$$\mathbf{Y}_l = (l_T/m_1)\ (\mathbf{Z}_l - \mathbf{1}_{n_l}\bar{\mathbf{z}})\,\mathbf{R}_T^{-1} + \mathbf{1}_{n_l}\bar{\mathbf{x}}_T$$
$$\mathbf{X}_l = (l_T/m_1)\ (v/v_T)^{1/3}\ (\mathbf{Z}_l - \mathbf{1}_{n_l}\bar{\mathbf{z}})\,\mathbf{R}^{-1} + \mathbf{1}_{n_l}\bar{\mathbf{x}}$$

, where the landmarks in the atlas and individual pigs are denoted $\mathbf{Y}_l$ and $\mathbf{X}_l$, respectively.

2.6 Non–rigid transformation

The stacked matrices of $\mathbf{Y}_l$–s and $\mathbf{X}_l$–s (all bones), are denoted $\mathbf{Y}_1$ and $\mathbf{X}_1$. These matrices were used to construct a cubic B-spline based transformation of $\mathbf{X}_1$ to $\mathbf{Y}_1$. The underlying model for the transformation is:

$$\mathbf{Y}_1 = \mathbf{Q}_{1X}\beta_1 + \mathbf{E}_1,$$

, where $\mathbf{Q}_{1X}$ denotes a matrix of size $n_1 \times p_L$ the elements of which were calculated by tensor (cubic) B-spline functions using $\mathbf{X}_1$ as input. The parameter β_1 denotes the regression parameters and $\mathbf{E}_1$ random noise. We utilized existing software (Kroon, 2011a,b) for the implementation of all B-spline based transformations. The software automatically calculated $\mathbf{Q}_{1X}$ including optimizing the knot grid used in the cubic B-spline functions, and provided estimates, $\hat{\beta}_1$, of β_1 for all pigs based on the input $\mathbf{X}_1$ and $\mathbf{Y}_1$.

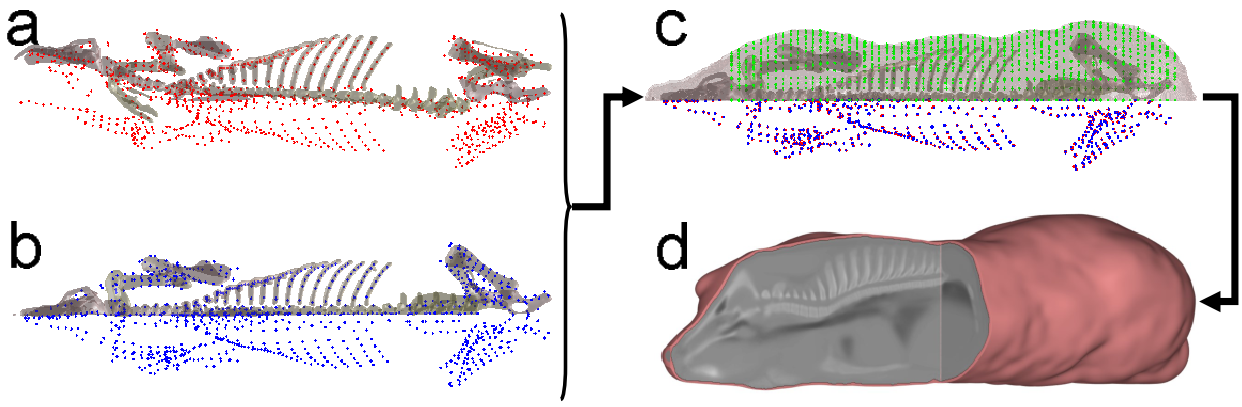


Figure 2: Construction of corresponding landmarks and the intensity atlas. (a) Landmarks for all bones transformed back to the original space of the pig. (b) Landmarks of all average bones transformed to the atlas space. (c) Non–rigid transformation based on the skeleton landmarks applied to the skeleton (blue/ red) and surface (skin). A secondary set of landmarks on the pig surfaces (green). (d) The intensity atlas. I.e. average HU–units after all voxels of all pigs are transformed to the atlas space.

For all pigs the surface voxels (skin) were identified, with coordinates denoted $\mathbf{X}_S$. The surface points from all 386 pigs were transformed to a common 3D image, $\mathbf{S}$, with the same dimensions as the atlas, by applying the transformation based on skeleton landmarks. The mathematical formula for this transformation is written as $\hat{\mathbf{Y}}_S = \mathbf{Q}_{SX}\hat{\beta}_1$ where the rounded values of $\hat{\mathbf{Y}}_S$ gave the coordinates of the surface voxels $\mathbf{X}_S$ transformed to $\mathbf{S}$. In order to get a symmetric surface, $\mathbf{S}$ was mirrored over the sagittal plane. The final atlas surface was defined as the voxels in $\mathbf{S}$ having maximum intensity and composing a continuous, connected surface.

For every 20 mm, on the interval from 200mm to 1400mm, along the longitudinal axis of the atlas surface, 34 new landmarks were set on the average surface (Fig. 2c). These points were set at a fixed set of angles around the centre of the slice in question. The coordinates of these landmarks are denoted $\mathbf{Y}_2$. Corresponding points for individual pigs, $\mathbf{X}_2$, were set as the surface points in $\mathbf{X}_S$ of which the corresponding transformed points, i.e. $\hat{\mathbf{Y}}_S$, had the minimum Euclidian distance to the points in $\mathbf{Y}_2$.

The motivation for constructing the corresponding points on the surface, i.e. $\mathbf{Y}_2$ and $\mathbf{X}_2$, was to increase the precision of the final B-spline transformations that were applied to the full volumes of the original pigs. Hence, the coordinates of the full volumes were the rounded values of $\hat{\mathbf{Y}} = \mathbf{Q}_{12X}\hat{\beta}_{12}$, where the basic functions of $\mathbf{Q}_{12X}$ and $\hat{\beta}_{12}$ were calculated using the stacked matrices of $\mathbf{Y}_1$ and $\mathbf{Y}_2$, and $\mathbf{X}_1$ and $\mathbf{X}_2$. The final intensity–based result is illustrated in Figure 2d. The intensities of the voxels in the intensity atlas are simply the average HU-unit after the final transformation of all voxels in all pigs.

2.7 Labelled atlas – atlas segmentation

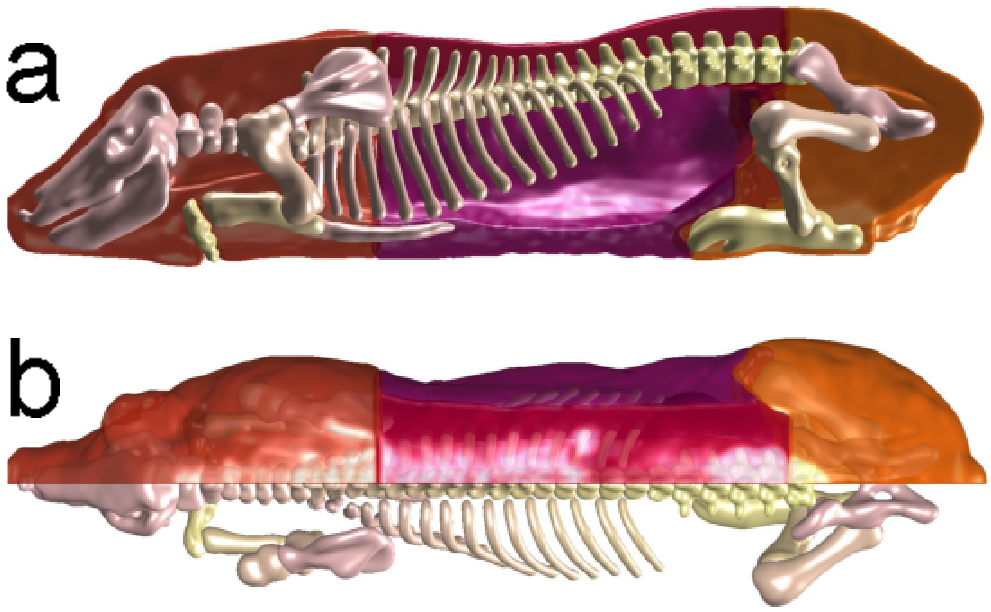


Figure 3: The labelled atlas. (a) View perpendicular to the sagittal plane. (b) View perpendicular to the coronal plane. In both panels ham is shown with orange color, belly with violet color, loin with clear red color and shoulder with red/ brown color. The major bones in the skeleton are shown withe different shades in gray/ yellow/ pink colors.

A labelled version of the atlas (Fig. 3a–b), was constructed by manual segmentation of the intensity atlas. The final step was to transform the labels onto the individual pigs, or eventually, onto new pigs registered to the atlas. Since every voxel in the individuals transformed to the (labelled) atlas corresponds to exactly one voxel in the atlas, the label of all voxels in individual pigs are easily defined (Fig. 4a–d).

The inner organs were segmented out by methods combining thresholds (HU-units) in the intensity atlas, and manual segmentation. The commercial cuts were set by segmenting shoulder from loin and belly by a cut exactly in the transverse plane of the atlas. The ham and loin were also segmented by a cut in the transverse plane. Belly was segmented from ham and loin by manual segmentation based on the intensity atlas.

2.8 Validation

First and foremost the method was validated by visual inspection of the segmentation applied to the individual pigs.

In order to conduct a numerical test of the method, we applied atlas segmentation to the primal cuts of 52 carcasses (left half) (Fig. 4). We predicted the weights of all voxels by applying a simple regression equation for voxel density (kg/m^3) using the intensities,

measured as Hounsfield units (HU), as predictor variable. The regression parameters were calculated by ordinary least squares regression using the registered weights of all 52 carcasses as response.

The corresponding cut weights (kg) and their proportions (% of carcass weight) (carcass right half) were registered by butchers at the Norwegian Meat and Poultry Research Center (Animalia) pilot plant. Thus we were able to calculate the correlations between cut weights and cut proportions based on two independent methods, i.e. atlas segmentation and manual butchering. Variances in cut proportions are, unlike variances in the cut weights, independent of total carcass weight. Thus, an eventual significant positive correlation for cut proportions, as opposed to the correlation between cut weights, might be viewed as a strong indication of the validity of the atlas segmentation method.

2.9 Code availability

All computations were conducted using the software Matlab. A demonstration of the central parts of the computer code applied to data from parts of a random pig is included as supplementary material in the zipped folder "Code_and_Data.zip".

3 Results

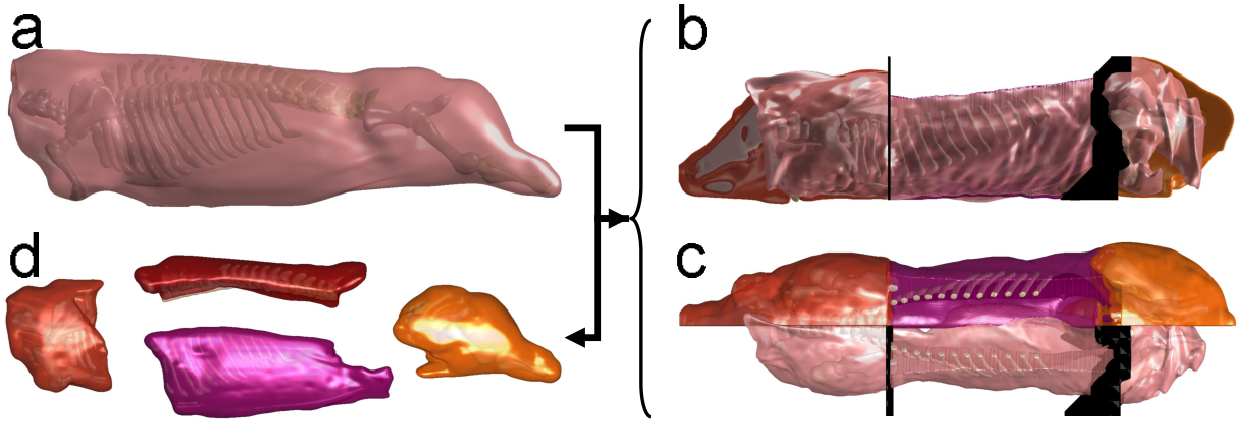


Figure 4: Atlas segmentation applied to a carcass (left half). (a) An untransformed carcass. (b–c) The carcass registered (transformed) to the atlas. The loin cut is removed to increase visibility. The other cuts are illustrated as black surfaces. (d) The final segmentation for the carcass in its four major cuts.

Visual inspection of the individual carcasses after transformation show that the method has an acceptable accuracy for atlas segmentation of the major parts, for an example see supplementary Video 1. The accuracy is best close to the skeleton structure, where the

density of landmark is huge, whereas the accuracy declines in areas where landmarks are scarce, typically in the back part of the belly.

The correlations between cut weight measured by atlas segmentation and manual butchering were 0.95, 0.91, 0.87 and 0.95 for shoulder-, belly-, loin- and ham weights, respectively. For the cut proportions the corresponding correlations were 0.60, 0.38, 0.36 and 0.47, all significantly different from 0 ($p < 0.01$). The variation in cut proportions between individuals were small, i.e. standard deviation at approximately 1 % unit.

4 Discussion

Differences in predicted cut weights between left and right sides might be substantial due to morphological differences, butcher effects and inaccurate splitting of carcasses. For shoulder and belly weights, differences between butchers are reported as high as 6–10% (Nissen et al., 2006). Thus, the correlation between the cut weights registered by butchers and by atlas segmentation was not expected to be extremely high even with a perfect atlas segmentation. For the cut proportions the *a priori* expected correlation between the two methods were substantially lower, due to the small variation in cut proportions between individuals. Thus, the highly significant positive correlations is a strong support for the usefulness of atlas segmentation.

The transformations were solely based on corresponding landmarks. The state–of–the–art methods in medical image analysis would generally include an extra step involving fine tuning of the transformation based on image intensities, typically based on the Gauss–Newton algorithm (Gill and Murray, 1978). This step aims at minimizing the cost based on a similarity measure between individual pigs and the intensity atlas (reference and template), utilizing the intensities of all data points. The transformations and intensity atlas described in this paper would constitute a natural starting point for such an algorithm. If successful, the result would be an even finer tuned intensity atlas, which in turn enables construction of a more detailed labelled atlas. However, there is a substantial risk associated with such methods as they may result in convergence to local optima, or yield over–fitted solutions, i.e. applying too much non-rigid deformation.

For a whole-body analysis, the corresponding landmarks are sufficient to obtain a satisfactory level of accuracy. As the method is automatic and robust, it offers a potential of multiplying the level of registered phenotypic variation for the full parental lines of breeding pigs. Thus it might constitute the foundation for the next generation of high-throughput and high-density phenotyping in animal breeding.

Acknowledgements

L.E.G. was supported by the Research Council of Norway, grant 225294 (PigComp).

Author contributions statement

L.E.G., J.K. wrote the main manuscript text. L.E.G., J.K. and S.S. analyzed the results and wrote computer code. J.K. and E.G. conceived the concept of implementation atlas segmentation to the breeding program. K.O. was involved in aspects involving pig anatomy. All authors reviewed the manuscript.

Additional information

Competing financial interests J.K. and E.G. are employed by Norsvin SA, who plan to implement the methods described in the paper in their pig breeding program. L.E.G. is offered employment partly financed by Norsvin SA via grant 256316 from the Research Council of Norway. K.O. and S.S. declare no competing financial interests.

References

Animal welfare Act 2009-06-19-97 (in Norwegian) (2009). Lov om dyrevelferd. Lovdata. (Accessed: 19th February 2016).

Baiker, M., Milles, J., Dijkstra, J., Henning, T. D., Weber, A. W., Que, I., Kaijzel, E. L., Löwik, C. W., Reiber, J. H., and Lelieveldt, B. P. (2010). Atlas-based whole-body segmentation of mice from low-contrast Micro-CT data. *Medical Image Analysis*, 14(6):723–737.

Commowick, O. (207). *Design and Use of Anatomical Atlases for Conformal Radiotherapy Planning*. PhD thesis, INRIA Sophia Antipolis, France.

Gangsei, L. E. and Kongsro, J. (2016). Automatic segmentation of Computed Tomography (CT) images of domestic pig skeleton using a 3D expansion of Dijkstras algorithm. *Computers and Electronics in Agriculture*, 121:191–194.

Gill, P. E. and Murray, W. (1978). Algorithms for the Solution of the Nonlinear Least-Squares Problem. *SIAM Journal on Numerical Analysis*, 15(5):977–992.

Gjerlaug-Enger, E., Kongsro, J., Ødegård, J., Aass, L., and Vangen, O. (2012). Genetic parameters between slaughter pig efficiency and growth rate of different body tissues estimated by computed tomography in live boars of Landrace and Duroc. *Animal*, 6(01):9–18.

Houle, D., Govindaraju, D. R., and Omholt, S. (2010). Phenomics: the next challenge. *Nature Reviews Genetics*, 11(12):855–866.

Hu, M.-K. (1962). Visual pattern recognition by moment invariants. *Information Theory, IRE Transactions on*, 8(2):179–187.

Kroon, D.-J. (2011a). B-spline Grid, Image and Point based Registration. MATLAB Central File Exchange. (c) 2009; Dirk–Jan Kroon. Retrieved June 02, 2015.

Kroon, D.-J. (2011b). *Segmentation of the Mandibular Canal in Cone-beam CT Data*. PhD thesis, University of Twente, Enschede, The Netherlands.

Li, X., Yankeelov, T. E., Peterson, T. E., Gore, J. C., and Dawant, B. M. (2008). Automatic nonrigid registration of whole body CT mice images. *Medical physics*, 35(4):1507–1520.

Nissen, P. M., Busk, H., Oksama, M., Seynaeve, M., Gispert, M., Walstra, P., Hansson, I., and Olsen, E. (2006). The estimated accuracy of the EU reference dissection method for pig carcass classification. *Meat science*, 73(1):22–28.

Regulation for the keeping of pigs in Norway 2003-02-18-175 (in Norwegian) (2003). For 2003-02-18 nr 175: Forskrift om hold av svin. Lovdata. (Accessed: 19th February 2016).