

On the statistical calibration of analytical fingerprints for rapid characterisation of complex hydrocarbon mixtures

Philip Jonathan
Shell Research

International Biometrics Society
British Region Meeting on Chemometrics
York
31st March 2006



Abstract

- Analytical fingerprinting techniques are widely used in many fields
- Rapid physico-chemical characterisation
- Challenges for analytical chemist and statistician in each area of application.
- Statistically, range of available calibration approaches increased in recent years.
- We'll consider spectroscopic fingerprinting of complex hydrocarbon mixtures within the oil refining industry
- We'll look at opportunities for extension of calibration ideas

Overview

- Motivation: Analytical fingerprinting in oil manufacturing
- Case study 1: characterisation of 36 complex hydrocarbon mixtures using different spectroscopic methods
- Case study 2: characterisation of 76 complex hydrocarbon mixtures using infra-red spectroscopy
- Modelling: Extension of calibration methods
- Concluding remarks

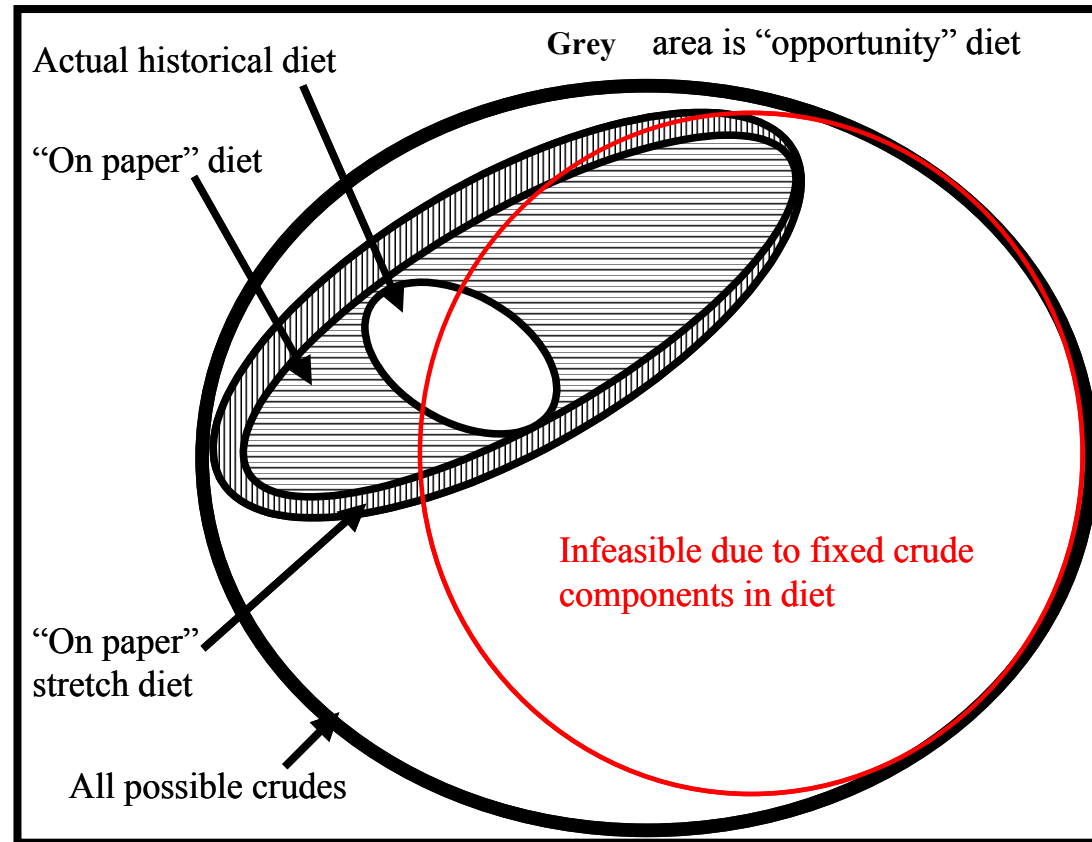
Analytical fingerprinting in oil refining

The oil business, “well to wheel”

- Find/grow/obtain fuel sources
 - predominantly complex hydrocarbon mixtures – crude oil
- Refine a wide variety of (mixtures of) different crude oils
 - widely varying hydrocarbon composition, impurities and contaminants
- Convert mixtures of low economic value to purer hydrocarbon mixtures with higher economic value
 - Complex network of processing steps, each performed in one or more process units.
- Basic chemical processes are
 - **distillation** - separation of hydrocarbon species by boiling point
 - **cracking** - conversion of very long-chained hydrocarbons to shorter chains
 - **alkylation** - conversion of short-chained hydrocarbons to longer chains
 - **impurity extraction** - e.g. of sulphur and nitrogen species, and trace metals
- End use (fuel heating / transportation, petrochemicals,...)

Opportunities for analytical fingerprinting

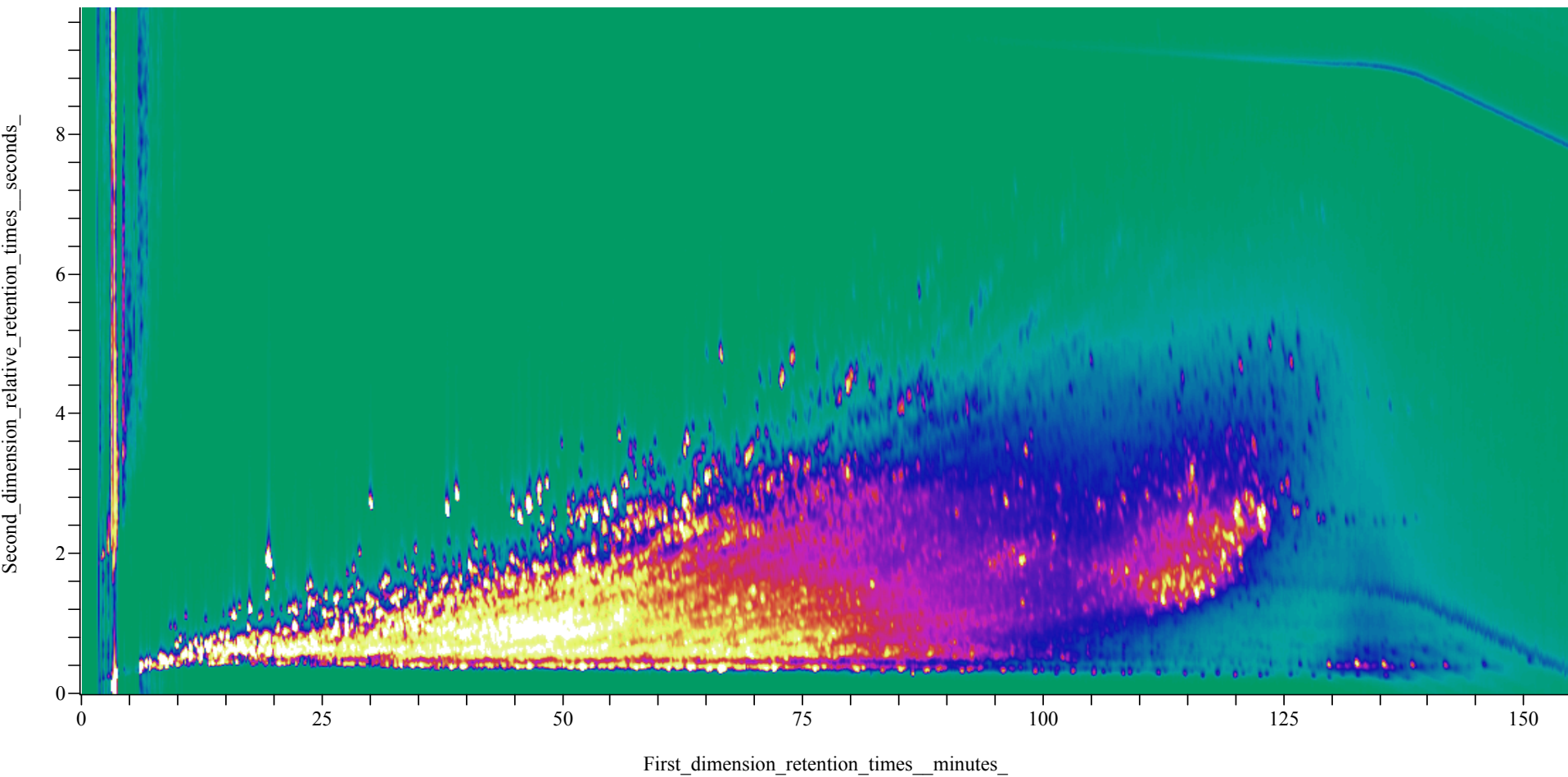
- General use of GC and NIR for characterisation of flows of hydrocarbon mixtures
- Concerns about long-term availability of crude from conventional sources
- Opportunities for biofuels, unconventional crudes, ...
- Need to know distillation curve, impurity levels
- Better "process control" in refining
- Requirement for better "molecular management" throughout the exploration, production, refining, retail.



Analytical fingerprinting in oil refining

A crude oil as seen by GC x GC

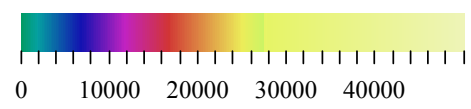
4/6



X="boiling point"

Y="polarity"

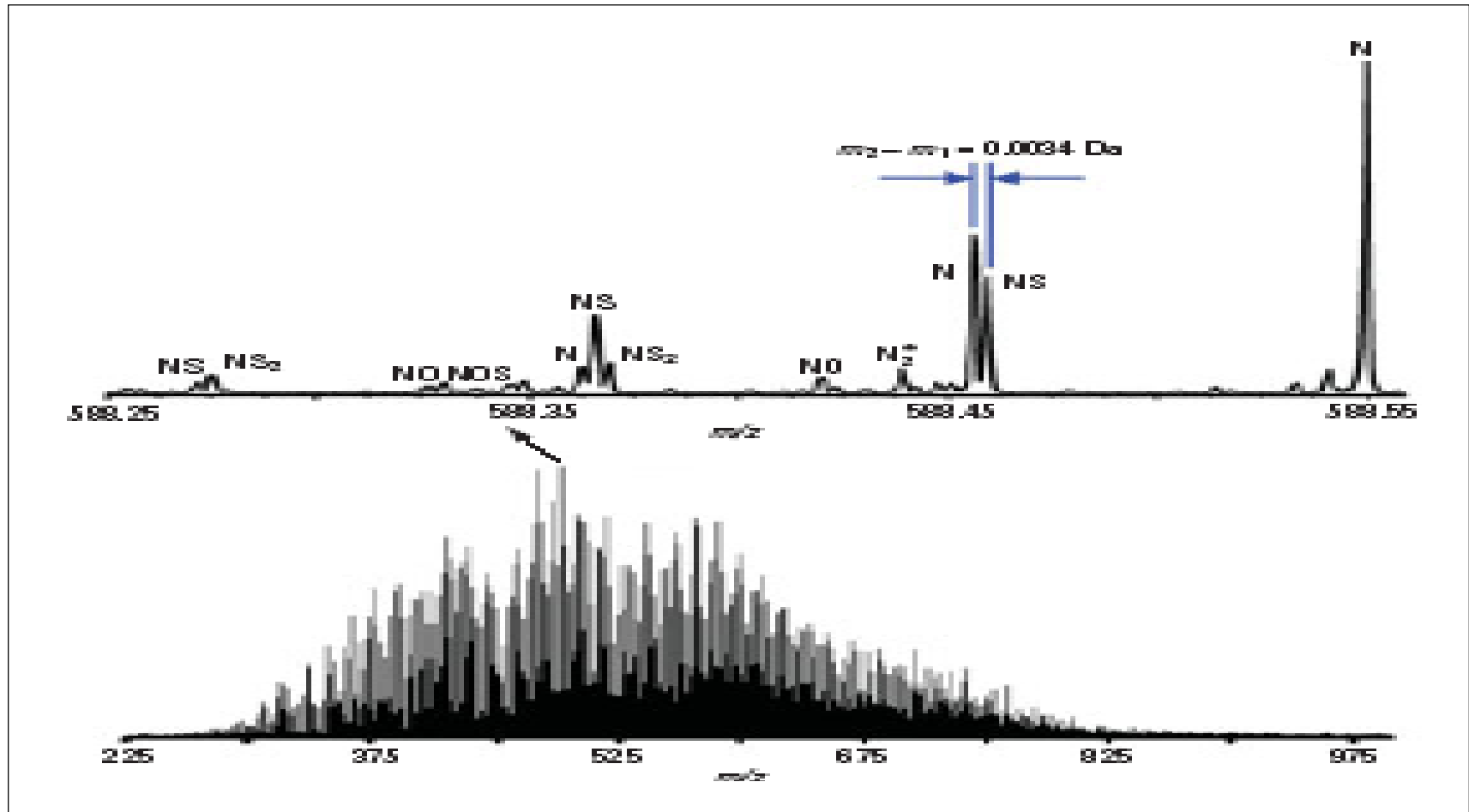
First_dimension_retention_times__minutes_



_2d2501z18_s_transform_txt_s

A crude oil as seen by FT ICR MS

(from Rodgers et al, Anal. Chem. 2005)



“Characterising 36”

Sample preparation and spectra acquisition

- Sample preparation:
 - Most samples viscous liquids
 - Some semi-solid or solid at room temperature
 - Heated to 40°C for 3 hours, homogenised and sub-sampled
 - Necessary to re-heat some samples for analysis
- Spectral acquisition
 - Difficult to establish set of experimental conditions suitable for all samples
 - Requirements for specific hardware enhancements (e.g a heated flow-through cell)

Characterising 36 Experimental design

Notes

- "Standard" measured multiple times throughout analysis
- "Standard" measured additionally at the start and end of each session of measurements
- Other individuals measured in random order
- "Light" and "Heavy" individuals measured twice (to gain some appreciation for measurement variability for "extreme" samples)

Run Order	NIR	IR/HAIR	H-NMR	C-NMR	Stdlist	UV	ICP-AES dilution	Anions	TAN-IP	TAN-SMS
1	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD
2	Q15	Q14	Q18	Q11	Q16	Q16	Q19	Q12	Q12	Q12
3	Q12	Q11	Q125	Q14	Q19	Q15	LIGHT	Q12	Q10	Q11
4	Q15	Q12	Q11	Q10	Q10	Q114	Q121	Q110	Q18	Q18
5	Q16	HEAVY	Q13	Q16	Q18	Q121	Q113	Q114	Q16	LIGHT
6	Q18	Q10	Q127	Q121	Q15	Q122	Q124	Q19	Q114	Q15
7	Q14	Q18	Q12	Q17	Q16	Q12	Q119	Q125	Q15	Q12
8	Q12	Q111	Q122	Q114	Q13	Q127	HEAVY	LIGHT	Q10	Q16
9	Q18	LIGHT	Q123	Q125	Q18	Q118	Q110	Q123	Q13	HEAVY
10	Q14	Q13	HEAVY	HEAVY	Q122	Q120	Q16	HEAVY	Q17	Q15
11	LIGHT	Q16	Q10	Q11	Q19	Q115	Q18	Q127	LIGHT	Q123
12	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD
13	Q10	Q18	Q121	Q122	Q11	Q11	Q114	Q116	Q127	Q10
14	Q17	Q115	Q126	Q127	LIGHT	HEAVY	Q112	Q111	Q110	Q18
15	Q13	HEAVY	Q114	Q117	Q124	Q119	Q126	Q113	Q19	Q19
16	Q12	Q17	Q16	LIGHT	Q111	Q17	Q123	LIGHT	Q18	Q17
17	Q126	Q15	Q115	Q113	Q114	Q124	Q12	Q126	Q112	Q13
18	Q16	Q121	Q116	Q126	Q120	Q14	Q117	Q17	Q13	Q11
19	Q19	Q19	HEAVY	Q10	HEAVY	Q19	Q111	Q121	Q14	Q111
20	HEAVY	LIGHT	Q124	Q15	Q126	Q116	Q11	Q119	Q19	Q13
21	Q10	Q127	LIGHT	Q18	Q15	Q110	Q15	Q18	Q121	Q16
22	Q19	Q14	Q117	Q119	Q117	Q126	Q17	Q117	Q14	Q121
23	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD
24	Q125	Q12	Q19	LIGHT	Q110	Q125	HEAVY	Q118	Q119	Q127
25	Q127	Q13	Q110	Q12	Q12	Q117	Q115	HEAVY	Q117	Q114
26	Q11	Q124	Q119	Q115	LIGHT	HEAVY	Q116	Q14	Q115	Q14
27	Q13	Q110	Q112	Q124	Q125	Q10	Q14	Q124	Q11	Q10
28	Q13	Q126	LIGHT	Q123	Q17	Q112	Q118	Q15	HEAVY	Q110
29	LIGHT	Q123	Q17	Q10	Q121	Q111	Q120	Q112	Q17	Q126
30	Q121	Q112	Q15	HEAVY	Q14	LIGHT	Q127	Q120	Q10	HEAVY
31	Q111	Q119	Q14	Q18	HEAVY	Q113	Q122	Q11	Q12	Q119
32	HEAVY	Q116	Q13	Q116	Q112	LIGHT	LIGHT	Q13	Q16	Q124
33	Q117	Q117	Q18	Q19	Q127	Q18	Q125	Q115	HEAVY	Q17
34	Q114	Q125	Q111	Q112	Q123	Q123	Q13	Q16	STANDARD	STANDARD
35	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	Q15	Q12
36	STANDARD	STANDARD		STANDARD	STANDARD	STANDARD	STANDARD	STANDARD	Q18	Q125
37	Q15	Q18		Q18	Q15	Q12	Q13	Q17	Q11	Q14
38	Q14	Q14		Q15	Q17	Q17	Q12	Q15	Q12	Q18
39	Q11	Q10		Q16	Q11	Q14	Q14	Q19	Q111	Q15
40	Q10	Q13		Q13	Q13	Q10	Q17	Q16	Q14	Q16
41	Q16	Q15		Q12	Q12	Q16	Q10	Q18	LIGHT	Q19
42	Q18	Q19		Q17	Q16	Q15	Q19	Q14	Q126	Q112
43	Q12	Q16		Q10	Q14	Q11	Q11	Q13	Q125	LIGHT
44	Q17	Q11		Q19	Q10	Q13	Q15	Q12	Q12	Q17
45	Q13	Q17		Q14	Q18	Q19	Q18	Q11	Q116	Q10
46	Q19	Q12		Q11	Q19	Q18	Q16	Q10	STANDARD	STANDARD
47	STANDARD	STANDARD		STANDARD	STANDARD	STANDARD	STANDARD	STANDARD		
		IR Transmission					ICP-AES using			
		LIGHT					LIGHT			
		STANDARD					STANDARD			
		HEAVY					HEAVY			

Spectral registration: Literature

- General: Ramsay & Silverman: Functional data analysis
 - “shift” registration (“time-warping”)
 - “feature” or “landmark” registration
 - warning that the registration process can remove real “signal”
- Specific pre-processing of spectra, e.g.
 - multiple scatter correction for NIR (e.g. Geladi et al (1985) *AppSpec* 39 491)
 - Vogels et al (TNO, 1996) : “Partial linear fit: a new *NMR* spectroscopy preprocessing tool for pattern recognition applications” *JChemo* 10 425
- Effects of instrument perturbations, e.g.
 - Estienne et al (Brussels, 2004) *ChemoLab* 73 207

Spectral registration: 36 set

- Multiple scatter correction for NIR
- Baseline correction (linear or cubic spline)
- Elimination of solvent peaks (NMR)
- Normalisation to unit total signal
- “Landmark registration” (NMR for 76)

Canonical correlation analysis

e.g. Mardia, Kent & Bibby (1979), Krzanowski (1988)

For samples in variables x and y , find those linear combinations $a'x$ and $b'y$ which have maximum possible correlation.

We might think of $a'x$ as best predictor, and $b'y$ as most predictable.

The solution is based on SVD of $S_{xx}^{-1/2} S_{xy} S_{yy}^{-1/2} = ULV'$.

The canonical correlation vectors are $S_{xx}^{-1/2}U$ and $S_{yy}^{-1/2}V$.

The square roots of the elements of (diagonal) L are the canonical correlation coefficients

Calibration using ridge regression

- Hoerl & Kennard (1970) *Techno* 12 55

Ridge regression seeks a solution to the multiple regression problem:

$$\underline{y} = \underline{X}\underline{\beta} \quad \text{so that} \quad \hat{\underline{\beta}} = \underset{\underline{\beta}}{\arg \min} \left[\|\underline{y} - \underline{X}\underline{\beta}\|^2 + \lambda \|\underline{\beta}\|^2 \right] = (\underline{X}'\underline{X} + \lambda \underline{I})^{-1} \underline{X}'\underline{y}$$

Cross-validation necessary to select the value of the shrinkage parameter λ which minimises the prediction error.

Many comparisons over the years of relative merits of ridge regression compared with PLS etc:

Frank & Friedman *Techno* 35 109

Forrester & Kalivas (2004) *JChemo* 18 372 discuss selection of the optimal ridge parameter taking into account both prediction error and variance effects, and find that PLS, PCR and RR perform similarly (in terms of prediction error and variance) given considered selection of tuning parameters. (c.f. Denham (2000) *JChemo* 14 351 for PLS)

Extending calibration methods

Overview

- Exploit serial correlation structure of spectra explicitly – regularisation methods
- Relax linearity constraints
- Select subset of variables efficiently
- Model validation, confidence, “interestingness”
- Model aggregation

Exploiting serial correlation structure using regularisation

Roughness penalty: $\arg \min_f \sum_i (y_i - f(x_i))^2 + \lambda \int_t [f''(t)]^2 dt$ given λ

f might be a spline

Impose our beliefs concerning form of solution (e.g. smoothness)

Basis for functional data regression

(Hastie, Tibshiriani & Friedman (2001) E.S.L)

Related ideas in chemometrics and related disciplines:

Friedman (1991) *AnnalsStats* 19 1 - MARS regression splines

Goutis (1998) *JRSSB* 60 103 - Second derivative functional regression

Marx & Eilers (2005) *JASA* 147 13 - Multivariate penalised signal regression

Yao & Lee (2006) *JRSSB* 68 3 - Penalised splines for functional PCA

No benefit in current application

Non-linear solutions: kernel methods

$K(x, z) = \langle \varphi(x), \varphi(z) \rangle$ where φ is a mapping to a feature space.

e.g. if $K(x, z) = \langle x, z \rangle^2 = x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1 x_2 z_1 z_2 = \langle (x_1^2, x_2^2, \sqrt{2}x_1 x_2), (z_1^2, z_2^2, \sqrt{2}z_1 z_2) \rangle$

we can view the corresponding φ as a mapping $\mathbb{R}^2 \rightarrow \mathbb{R}^3$

Kernel versions of PCA, ridge regression etc are easily produced, using the sample inner product matrix in feature space. Gaussian and polynomial kernels are popular.

Closely related to many ideas like nearest neighbours, support vector machines, regularisation.

Schölkopf & Smola (2002) - Learning with kernels

Czekaj et al (2005) *JChemo* 19 341 - Kernel ridge regression etc.

Radial basis kernel ridge regression provided slight improvement in current application.

Variable subset selection

- MCMC
 - Stochastic search variable selection:
 - George & McCulloch (1996) *McmcInPractice*
 - Bayesian wavelength selection:
 - Brown et al (1998) *JChemo* 12 173
- Variable selection with autocorrelated errors:
 - Brown (1993)
- LASSO & LARS
 - Regression shrinkage and selection via the lasso
 - Tibshirani (1996) *JRSSB* 58 267
 - Least angle regression
 - Efron et al (2004) *AnnStat* 32 307
 - “Elastic net” for regularisation and variable selection ($p \gg n$)
 - Zou & Hastie (2005) *JRSSB* 67 301

Model validation

- Cross-validation
 - The dangers of overfitting using leave-one-out
 - Estienne et al (2004) *Chemolab* 73 207
 - Appropriate leave-out sample size, m
 - Forrester & Kalivas (2004) *JChemo* 18 372
 - Nested cross-validation (e.g. "2-deep")
 - Stone (1974) *JRSSB* 36 111
 - Jonathan et al (2000) *Stats&Comp* 10 209
- Confidence
 - Bootstrapping for confidence intervals on regression and predictions
 - Wehrens et al (2000) *Chemolab* 54 35 (tutorial) *Clearly, bootstrap methods offer many advantages in the analysis of real-world data, where we never can be sure about the validity of our assumptions.*
- Interestingness
 - Randomised permutation testing
 - Krzanowski et al *AppStat* 44 101

$$m = n \left(1 - \frac{1}{\ln(n) - 1} \right)$$

Model aggregation

e.g. Hastie, Tibshirani & Friedman (2001)

Bagging: bootstrap aggregation

- Aggregate predictions from a number of predictors: $f_B(x) = \frac{1}{B} \sum_{i=1}^B \hat{f}_i^*(x)$
- Each individual predictor \hat{f}_i^* is obtained by model building on a bootstrap sample from the original training data.
- Bagging reduces prediction variance.

Bumping: bootstrapping wrt two different optimisation criteria

- Draw bootstrap samples, model-build using one optimisation criterion (say, least squares), retaining only predictors $\{\hat{f}_i^*\}_{i=1}^B$ with good performance.
- Re-evaluate the predictive performance wrt a different but desirable criterion (say, least absolute error). Select that predictor which gives best performance wrt this criterion also.
- Bumping seeks model variants that are in some senses preferable. In particular, bumping will eliminate the effects of a small number of outliers in the original training data.

Model aggregation

Stacking: combine predictors of different types to obtain an aggregate model with better predictive performance.

- Combine, say, regression vectors with different numbers of terms from a stepwise selection procedure, with a spline and a kernel-based model
- First, individual predictors $\{\hat{f}_i\}_{i=1}^C$ (which may not be available in closed form) are estimated for each of the different contributor models, together with leave-out cross-validated predictions $\left\{\left\{\hat{f}_i^{-j}(x_j)\right\}_{j=1}^N\right\}_{i=1}^C$ for each individual in the training data.
- Estimate the weight vector $\underline{w} = \{w_i\}_{i=1}^C$ with which to aggregate the contributor models to yield best predictive performance (assessed by cross-validation):

$$\underline{w}_{OPT} = \underset{\underline{w}}{\arg \min} \sum_{j=1}^N \left(y_j - \sum_{i=1}^C w_i \hat{f}_i^{-j}(x_j) \right)^2$$

- Solution is linear algebra, unless additional constraints like $\|\underline{w}\| = 1$ and $\{w_i \geq 0\}_{i=1}^C$ added.
- c.f Breiman & Friedman *JRSSB* 59 3 (Curds & Whey - predicting multivariate responses in multiple linear regression)

Concluding remarks

- Illustrated the role of chemometrics in the oil industry
- Analytical fingerprints able to characterise complex hydrocarbon mixtures relatively well
- Key steps:
 - sample presentation, spectral acquisition
 - spectral registration prior to calibration modelling
 - model validation
- Linear ridge regression does as well as more sophisticated techniques
 - parsimony & simplicity vs model complexity

Thank you

philip.jonathan@shell.com