# Contents

1 Publication 2

2 **Summary of Research Work** 4
   2.1 Wavelets in statistics 5
      2.1.1 Wavelet based empirical Bayes estimation for the uniform distribution 5
      2.1.2 Density estimation by wavelet-based reproducing kernels 6
      2.1.3 Bayesian wavelet shrinkage for nonparametric mixed effects models (joint with H.S. Lu) 6
      2.1.4 Extended Gauss–Markov Theorem for nonparametric mixed-effects models (joint with H.S. Lu) 8
      2.1.5 Generalized cross-validation for wavelet shrinkage in nonparametric mixed-effects models (joint with H.S. Lu and F.J. Lin) 9
      2.1.6 On a Bayesian aspect for soft wavelet shrinkage estimation under an asymmetric linex loss 10
      2.1.7 Wavelets, Advanced (joint with Z.D. Bai) 11
   2.2 Bayesian/empirical Bayesian inference 11
      2.2.1 Empirical Bayes estimation of the truncation parameter with Linex loss (joint with T. Liang) 11
      2.2.2 Bayesian marginal inference via Candidate’s formula (joint with C.K. Hsiao and C.W. Chang) 11
      2.2.3 Optimal volume-corrected Laplace-Metropolis method for computing marginal probability (joint with C.K. Hsiao and C.W. Chang) 12
   2.3 Miscellaneous 12
      2.3.1 A characterization of the uniform distribution via moments of n-fold convolution modulo one (joint with Y. Chow) 12
      2.3.2 Bernoulli numbers and polynomials via residues (joint with I-C. Huang) 13
      2.3.3 Kernel-based discriminant techniques on educational placement (joint with M.H. Lin and Y.C. Chang) 13

3 **Plan of Future Research** 14
   3.1 Continuing research 14
   3.2 Statistical learning theory 14

4 References 16
1 Publication


2 Summary of Research Work

My research work from 1997 to present can roughly be classified into three parts:

- **Wavelets in statistics.**
  - In article [6], which is about “wavelet based empirical Bayes estimation”, the wavelet technique is introduced into the nonparametric empirical Bayes method. Related asymptotic behavior is studied.
  - In article [8], which is about “wavelet based reproducing kernel estimator”, the asymptotic behavior of the linear wavelet estimator is studied via the framework of reproducing kernels Hilbert space. This article provides pointwise asymptotic bias and variance expressions for the linear wavelet estimator.
  - The BLUPWAVE sequel [11] [12] [15] is a study of wavelet shrinkage stemming from the best linear unbiased prediction (also known as Gauss-Markov prediction) for nonparametric mixed-effects models with Bayesian interpretation. The theoretical study of the Gauss-Markov prediction [12] helps in guiding the exploration for effective nonlinear estimates adaptively. This Gauss-Markov prediction based approach, implemented in the wavelet Sobolev regularization, results in a shrinkage scheme (we name it BLUPWAVE) [11] compromising between the classical soft and hard thresholds. There are parameters involved in the BLUPWAVE, which are selected by a data driven GCV procedure [15].

Article [12] is a general study on Gauss-Markov prediction. The idea behind this work is to extend the usual generalized (or weighted) least squares approach in finite dimension to some general functional space, say a certain $H_1$. The fixed effects and random effects in the finite dimensional LS approach transfer into a subspace decomposition in $H_1$. To define an inverse problem in $H_1$, some distance metric has to be adopted. Here we choose the norm of a reproducing kernel Hilbert space, which provides a natural extension of the LS norm in finite dimension to a functional space. Such a choice of functional norm enables us to obtain parallel but more general results as in the classical LS approach. Applications of this approach include deconvolution, wavelet Sobolev regularization [12], curve data analysis, image recovery, etc.

Article [11] is a study of wavelet shrinkage via (asymptotic) BLUP with emphasis on its Bayesian interpretation. The function or signal is decomposed into two components: the main feature (or low frequency part) is treated as fixed effects and the fine feature (or high frequency part) is treated as random effects (modelled via a Gaussian process, which is the prior in Bayesian language). This modelling treatment is a special case of the Gauss-Markov prediction discussed in [12]. It corresponds to a particular type of reproducing kernel Hilbert spaces, namely the Besov type. As for which one of the Besov spaces the prediction problem resides in (i.e., which one the prior space is), it is determined by the prior covariance kernel of the above mentioned Gaussian
process. In this article, we obtain properties concerning Besov space of prior sample paths. As for the resulting predictor (posterior function or posterior signal in Bayesian language), it falls into a smoother class (with regularity 1/2 higher) of function space of Sobolev type. In other words, through this Gauss-Markov prediction as a denoising scheme, the prior signal becomes smoother with a posterior regularity 1/2 higher.

In Article [15], a GCV method is used to select parameter values involved in the wavelet shrinkage [11]. The method can be used for both level independent as well as level dependent thresholdings. It can also be used to select the primary resolution level and the degree of smoothness in wavelet basis. Besides extensive simulation study, the article also contains some theoretical discussions on the strong consistency and assumptions for it to be valid. Though the GCV is a data-driven approach, the theoretical study provides some insight toward an “ideal” threshold for the BLUPWAVE shrinkage. Both the soft and BLUPWAVE consist of a keep-or-kill step and a shrinkage step. From a theoretical viewpoint, an ideal BLUPWAVE threshold tends to be more conservative (keep more nonzero coefficients) in the keep-or-kill step than the soft shrinkage (with universal threshold or GCV selection alike), while more drastic in the shrinkage step.

– In the “Bayesian aspect for soft wavelet shrinkage” article [13], we render the soft wavelet shrinkage a Bayesian meaning.

– Article [14] is a review for the ENCYCLOPEDIA OF PHYSICAL SCIENCE & TECHNOLOGY by Academic Press.


• Miscellaneous. There are 3 articles. One is on a characterization of uniform distribution based on a class of identities involving Bernoulli polynomials of n-fold convolution modulo one [9]. One is on Bernoulli numbers and polynomials represented as residues of local cohomology classes and some related results [10]. The other is a statistical application on educational placement [18].

Listed below are itemized summaries in chronic order within category for my publications.

2.1 Wavelets in statistics

2.1.1 Wavelet based empirical Bayes estimation for the uniform distribution

In this article, we incorporate the wavelet tool into the problem of empirical Bayes estimation. The asymptotic behavior of the wavelet based empirical Bayes estimator is investigated. The kernel based estimator studied by Nogami (1988) has convergence rate $O(n^{-1/2})$. We show that the proposed wavelet based empirical Bayes estimator attains the rate $O(n^{-2s/(2s+1)})$, where $s \geq 1$ is the regularity index of the marginal pdf.
2.1.2 Density estimation by wavelet-based reproducing kernels

In wavelet-based function estimation, various nonlinear shrinkage estimators are known to be effective in recovering noise-contaminated functions and can be well adapted to local features of functions of interest. Such nonlinear wavelet shrinkage estimators involve a linear component and a nonlinear component. The linear part, consisting of the primary resolution level, corresponds to the main feature of the underlying function. In this article we look into the asymptotic behavior of the linear part, the so-called linear wavelet estimator.

The linear wavelet estimator is studied through the viewpoint of reproducing kernels, which are different from the convolution type kernels. Concepts such as kernel order and kernel symmetry are defined and discussed and linked to the multiresolution approximation. Explicit expressions of pointwise asymptotic bias and variance are derived. Examples using reproducing kernels from spline-wavelets and Daubechies wavelets are presented. A bias reduction technique based on a grid point average is proposed and shown to be variance stable.

2.1.3 Bayesian wavelet shrinkage for nonparametric mixed effects models (joint with H.S. Lu)

We study the wavelet shrinkage estimation from a Bayesian viewpoint. Nonparametric mixed-effects models are proposed and used for interpretation of the Bayesian structure. Under the nonparametric mixed-effects models the best linear unbiased prediction (BLUP) is derived. An asymptotically equivalent expression of the BLUP is derived, which involves hyperparameters. The hyperparameters are then estimated from the empirical data. We name the resulting nonlinear estimator BLUPWAVE. The BLUPWAVE combines the asymptotic equivalence to the best linear unbiased prediction and the Bayesian estimation. Such an estimator has shrinkage effect compromising between hard and soft thresholds by drawing positive aspects from both strategies. The Bayesian characterization of prior and posterior spaces is discussed. The smoothness of posterior estimators is controlled via hyperparameters. Computational issues are discussed. A simulation is carried out for comparison with the Bayesian method by Abramovich, Sapatinas and Silverman (1998).

The Bayesian formulation in this article is conceptually inspired by the work of Parzen (1961), as well as by the work of Kimeldorf and Wahba (1970, 1971), and Wahba (1978, 1990). The equivalence to regularization and the minimaxity property for Bayesian wavelet shrinkage is influenced by the work of Li (1982). Our Bayesian setup and viewpoints are different from those in Vidakovic (1998), Clyde, Parmigiani and Vidakovic (1998), Chipman, Kolaczyk and McCulloch (1997), and Abramovich, Sapatinas and Silverman (1998). The setup here has some common structure with spline models and our modelling makes it easy to incorporate prior information on smoothness, to obtain optimality results, and to relate them to the method of regularization.

In the nonparametric regression problem, we have a response variable $y$ and a predictor variable $x$ linked by $y = f(x) + \epsilon$, where the noise $\epsilon$ is assumed to follow a normal distribution
\( N(0, \sigma^2) \). The aim is to estimate the mean regression function \( f \) based on independent observations \( y_i \) collected at \( x_i, i = 1, \ldots, n \). Expand the regression function \( f \) in terms of wavelet basis as

\[
 f(x) = \sum_k c_{j,k} \phi_{j,k}(x) + \sum_{\ell=j}^{\infty} \sum_k d_{\ell,k} \psi_{\ell,k}(x),
\]

(1)

where \( j \) is the primary resolution level, and \( \phi_s \) and \( \psi_s \) are scaling functions and wavelet functions respectively. For \( n = 2^J \) and \( \{x_i\} \) uniformly distributed, the empirical wavelet coefficients can be obtained by the discrete wavelet transform:

\[
w = \left( \frac{\sqrt{n} \hat{c}}{\sqrt{n} \hat{d}} \right) = Wy
\]

(2)

where \( y = (y_1, \ldots, y_n)^T \) and \( W \) is the orthogonal matrix associated with the DWT. To suppress the noise level, the empirical wavelet coefficients are put through a denoising process prior to the reconstruction of the target function. Each individual wavelet coefficient \( \hat{d}_{\ell,k} \) is shrunk towards zero via either a level-independent shrinkage scheme

level-independent BLUPWAVE: \( \Delta(\hat{d}_{\ell,k}, \delta) = \left( 1 - \frac{\delta^2}{\hat{d}_{\ell,k}^2} \right)_+ \hat{d}_{\ell,k}, \)

(3)

or a level-dependent shrinkage scheme

level-dependent BLUPWAVE: \( \Delta(\hat{d}_{\ell,k}, \delta_{\ell}) = \left( 1 - \frac{\delta_{\ell}^2}{\hat{d}_{\ell,k}^2} \right)_+ \hat{d}_{\ell,k}. \)

(4)

The BLUPWAVE shrinkage rule compromises between hard and soft thresholding by drawing positive aspects of both strategies. The shrinkage curve of BLUPWAVE falls between those of hard and soft thresholdings, as seen in Figure 1.

Often \( \delta \) or \( \delta_{\ell} \) are not known apriori, they are selected by GCV procedure. When compared with BayesThresh of Abramovich, Sapatinas and Silverman (1998) (wherein \( n = 1024 \)), our results are often better (having smaller average square errors). See Table 2 for comparison. The symbol ‘+’ means that our results are better, a blank means about the same, and ‘-’ means our results are worse. Our parameters selection is data-driven, while their parameters are hand-pick.

<table>
<thead>
<tr>
<th>SNR</th>
<th>( n )</th>
<th>Blocks</th>
<th>Bumps</th>
<th>HeaviSine</th>
<th>Doppler</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>1024</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>7</td>
<td>1024</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>1024</td>
<td>+</td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>1024</td>
<td>-</td>
<td>+</td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>
Some theoretical aspects of the proposed Bayesian approach is also discussed in this article, including properties of asymptotic BLUP, equivalent regularization, linear minimaxity, and Bayesian prior and posterior function spaces.

2.1.4 Extended Gauss–Markov Theorem for nonparametric mixed–effects models (joint with H.S. Lu)

Consider the estimation problem in nonparametric mixed-effects models. In this article the Gauss-Markov theorem for linear model is extended to nonparametric mixed-effects models. Such extended Gauss-Markov estimation (or prediction) is shown to be equivalent to a nonparametric regularization method and its minimaxity is addressed. The resulting Gauss-Markov estimation serves as an oracle to guide the exploration for effective nonlinear adaptive estimators. Various examples are discussed. Particularly, the wavelet nonparametric regression example and its connection with a Sobolev regularization is presented.

Consider a process $Y$ observed through the model

$$Y = Af + \sigma \epsilon.$$  

(5)

The function of interest, $f$, is defined on an index set $T$. The index set $T$ can be an interval, a set of finite elements, or a set of countable elements. The error process $\epsilon$ is zero mean and is defined on an index set $J$ with a known covariance kernel $R$. $A$ is a linear mapping from $L_2(T)$ to $L_2(J)$. The underlying function $f$ is modelled via nonparametric mixed-effects. The function $f$ is expanded as

$$f(t) = \sum_{k=1}^{m} \beta_k \phi_k(t) + \delta Z(t) \text{ with } Z(t) = \sum_{\nu} \gamma_{\nu} \psi_{\nu}(t),$$  

(6)

where $\phi_k(t)$ and $\psi_{\nu}(t)$ are known functions constituting a complete basis for a certain function space, for instance, $L_2(T)$ or an approximate subspace of $L_2(T)$. The coefficients
\( \beta_k \) are fixed but unknown and the coefficients \( \gamma_\nu \) are uncorrelated random variables with zero mean and variance \( \lambda_\nu \). That is, \( Z(t) \) is a zero mean process with the covariance kernel 
\[ E[Z(t)Z(s)] = \mathcal{W}(t, s) = \sum_\nu \lambda_\nu \psi_\nu(t) \psi_\nu(s). \]
We also assume that \( Z(t) \) is independent of \( \epsilon \). The ratio \( \delta/\sigma \) is assumed known. Knowledge of the ratio \( \delta/\sigma \) together with the identifiability condition for \( \beta \) will be necessary and sufficient for the identifiability of the nonparametric mixed-effects model.

The mixed-effects model described in (5) and (6) is often used in the analysis of longitudinal data or curve data. (See Laird and Ware, 1982; Ramsay, 1982; Besse and Ramsay, 1986; Ramsay and Dalzell, 1991; Anderson and Jones, 1995; Ramsay and Silverman, 1997; among many others.) They also appear in the literature of spline smoothing and nonparametric Bayesian regression (Kimeldorf and Wahba, 1970 and 1971; Wahba, 1978 and 1990; and Barry, 1986). In this article we extend the Gauss-Markov Theorem for linear mixed-effects models (Harville, 1976) to the nonparametric mixed-effects models. In contrast to using traditional linear algebra approach, we use some analysis tool involving reproducing kernel Hilbert spaces (RKHSs). We find it quite natural to describe the space of fixed effects and the space of random effects via subspaces of RKHS and the penalty in the associated regularization can also be naturally represented by a semi-norm of the RKHS. The technical development of extended Gauss-Markov Theorem provides a unified and illuminating perspective for constructing linear estimators or predictors for various models. There are often parameters involved in the extended Gauss-Markov Theorem. When there is no prior knowledge about these parameters, one needs to estimate them based upon the data. Is then the extended Gauss-Markov Theorem of no practical value? The answer is no. The extended Gauss-Markov Theorem provides an oracle. Although it results in a linear estimator which often involves with unknown parameters, this linear estimator can serve as a useful guidance for exploring efficient and adaptive nonlinear estimators. The reader is referred to Huang and Lu (2000) for an application of the extended Gauss-Markov Theorem to wavelet nonparametric regression. The extended Gauss-Markov estimation is the so called best linear unbiased prediction (BLUP) and can be linked to a regularization method via functional penalized least squares (PLS).

2.1.5 Generalized cross-validation for wavelet shrinkage in nonparametric mixed-effects models (joint with H.S. Lu and F.J. Lin)

The GCV method is used to select the threshold parameters for the wavelet shrinkage estimation. The method can be used for both level independent as well as level dependent thresholdings. It can also be used to select the primary resolution level and the degree of smoothness in wavelet basis. The GCV method has low computational cost. Its strong consistency is proved in this article. Extensive simulation study is carried out for comparison among various wavelet shrinkage estimators.
2.1.6 On a Bayesian aspect for soft wavelet shrinkage estimation under an asymmetric linex loss

Consider an asymmetric Linex loss. We provide the soft wavelet shrinkage estimation a Bayesian interpretation under such loss.

Assume the following discrete noisy signal model obtained from a discrete wavelet transform:

\[ w = \theta + \epsilon, \]  

where \( w = (w_1, \cdots, w_n)^T \) are empirical wavelet coefficients, \( \epsilon = (\epsilon_1, \cdots, \epsilon_n)^T \) are iid normal random errors with zero mean and variance \( \sigma^2 \), and \( \theta = (\theta_1, \cdots, \theta_n)^T \) are the true wavelet coefficients. Let \( \delta^{soft}(w) \) be the soft wavelet shrinkage estimator by Donoho and Johnstone (1994) given by

\[
\delta^{soft}_i(w, \lambda) = \text{sign}(w_i)(|w_i| - \lambda)_+, \quad i = 1, \ldots, n, \tag{7}
\]

where \( \lambda > 0 \) is the threshold parameter.

Treat the above wavelet estimation problem as estimation for the mean vector \( \theta \) of a multivariate normal distribution \( w|\theta \sim N(\theta, \sigma^2 I) \). We employ an asymmetric Linex loss function (Varian, 1975 and Zellner, 1986) as error criterion:

\[
L(\theta, \delta(w)) = \frac{1}{n} \sum_{i=1}^{n} \left( e^{a_i\{\delta_i(w) - \theta_i\}} - a_i\{\delta_i(w) - \theta_i\} - 1 \right), \quad a_i \neq 0. \tag{8}
\]

Under such a loss function (8), we derive the generalized Bayes estimator with respect to the flat improper prior on \( R^n \):

\[
\delta^{GB}_i(w) = w_i - \frac{a_i \sigma^2}{2}, \quad i = 1, \ldots, n. \tag{9}
\]

Moreover, \( \delta^{GB}(w) \) is shown to be the unique admissible and minimax estimator.

For application of the Linex loss to wavelet estimation problem, we consider specifically the Linex loss with \( a_i \) values depending on signs of \( \theta_i \)

\[
a_i = \begin{cases} 
    c & \text{for } \theta_i \geq 0, \quad i = 1, \ldots, n, \\
    -c & \text{for } \theta_i < 0, \quad i = 1, \ldots, n,
\end{cases} \tag{10}
\]

where \( c > 0 \) is some constant. Such an error criterion discourages estimators from over-estimation in magnitude (i.e. in absolute value) and results in shrinkage estimation towards zero. Under such a loss criterion, the unique admissible and minimax estimator is given by

\[
\delta^{GB}_i(w) = w_i - \text{sign}(\theta_i) \lambda, \quad \text{where } \lambda = \frac{c \sigma^2}{2}.
\]

Usually the signs of parameters \( \theta_i \) are not known, a natural approach is to use \( \text{sign}(w_i) \) to estimate \( \text{sign}(\theta_i) \) and make truncation at zero. We then have the following empirical version of \( \delta^{GB} \)

\[
\delta^{soft}_i(w) = \begin{cases} 
    (w_i - \lambda) \vee 0 & w_i \geq 0 \\
    (w_i + \lambda) \wedge 0 & w_i < 0
\end{cases}
\]

\[
= \text{sign}(w_i)(|w_i| - \lambda)_+,
\]

which is the soft wavelet shrinkage estimator.
2.1.7 Wavelets, Advanced (joint with Z.D. Bai)

This is a review article for the Encyclopedia of Physical Science & Technology by Academic Press. It includes introduction to wavelets, continuous wavelet transform, discrete wavelet transform, multiresolution analysis, cascade algorithm, filters, examples of wavelets, wavelet packets, best basis selection, wavelets in statistics, wavelet thresholding and shrinkage, wavelet denoising, wavelet regularization and image processing.

2.2 Bayesian/empirical Bayesian inference

2.2.1 Empirical Bayes estimation of the truncation parameter with Linex loss (joint with T. Liang)

This article deals with the empirical Bayes estimation of the truncation parameter of the truncated family under the Linex loss. Nonparametric empirical Bayes estimator is proposed and its asymptotic optimality and convergence rate are investigated. Under certain mild conditions without any differentiability assumption on either the prior or the marginal distribution, it is shown that the proposed empirical Bayes estimator is asymptotically optimal with convergence rate of order $O(n^{-2/3})$, where $n$ is the number of past observations at hand. Simulation results on the performance of the proposed empirical Bayes estimator are also presented.

2.2.2 Bayesian marginal inference via Candidate’s formula (joint with C.K. Hsiao and C.W. Chang)

In the context of Bayesian inference, a nonparametric kernel estimate via Candidate’s formula is developed for computing the marginal density of the sample data. The estimate is computed based on Markov chains outputs. The deficiency of high dimensional density estimation, known as the curse of dimensionality, can be handled well in this particular Bayesian marginal inference problem. As the Candidate’s formula is valid for all parameter values in the prior support, the nonparametric Candidate’s estimate (which involves kernel density estimate using simulated posterior sample) can be evaluated at various prior points and take average. Such an approach makes use of more posterior sample points and hence can ease the tension caused by data sparseness. This nonparametric Candidate’s estimate does not require knowledge of full conditional densities. We find it a convenient and comprehensible way to estimate the posterior density. In this article the asymptotic behavior of the estimate is studied and the best point for evaluating the estimate is also derived. A simulation study is carried out for comparison with Laplace type estimates.
2.2.3 Optimal volume-corrected Laplace-Metropolis method for computing marginal probability (joint with C.K. Hsiao and C.W. Chang)

Computing marginal probability has been an important problem and common practice in Bayesian statistical inference. A renowned asymptotic approximation for the marginal probability is Laplace’s method. DiCiccio, Kass, Raftery and Wasserman (1997) consider a simulation-aided Laplace’s method via Markov chain Monte Carlo techniques, known as Laplace-Metropolis method. They introduce a local volume correction around the mode of integrand to improve the approximation and have suggested, based on their simulation experience, a correction volume of 5% probability. In this paper we provide a refinement for the default value of \( \alpha = 0.05 \). Based on an asymptotic analysis, we derive the theoretical optimal correction volume. The optimal choice is shown invariant under linear transformation of posterior variables, so that we can work on transformed posterior sample. We outline a simple implementation procedure for this optimal correction volume method and carry out a simulation study.

2.3 Miscellaneous

2.3.1 A characterization of the uniform distribution via moments of \( n \)-fold convolution modulo one (joint with Y. Chow)

In this note, we characterize the uniform distribution \( U(0,1) \) via a class of identities involving Bernoulli polynomials of \( n \)-fold convolution modulo one.

Let \( X_1, X_2, X_3, \ldots \) be a sequence of i.i.d. random variables on \([0,1)\) from a distribution \( F(x) \). In Lin (1988), the uniform distribution \( U(0,1) \) is neatly characterized by two moment conditions:

\[
E \max(X_1, X_2) = 2/3 \quad \text{and} \quad EX_1^2 = 1/3.
\]

Let the symbol \( \oplus \) be referred to addition modulo one. In the error analysis and bias reduction in a spline-based multi-resolution approximation, Huang (1999) observe that, when \( F \) is \( U(0,1) \), we have, for any \( d \in \mathbb{N} \),

\[
E B_d(X_1 \oplus X_2 \oplus \cdots \oplus X_n) = 0, \quad \text{for all} \ n \in \mathbb{N},
\]

where \( B_d(x) \) is the \( d \)-th Bernoulli polynomial defined recursively by

\[
B_0(x) = 1, \quad B'_d(x) = dB_{d-1}(x) \quad \text{and} \quad \int_0^1 B_d(x)dx = 0 \quad \text{for} \ d \geq 1.
\]

It is then natural to ask whether conditions of the type in (11) characterize the uniform distribution \( U(0,1) \). An affirmative answer is obtained under the technical assumption that

the density \( f(x) = F'(x) \) exists a.e. on \((0,1)\).
2.3.2 Bernoulli numbers and polynomials via residues (joint with I-C. Huang)

Bernoulli numbers and polynomials can be represented as residues of local cohomology classes. Computations on these local cohomology classes and residues yield new results involving Bernoulli numbers and polynomials. We are able to evaluate summations of the form

\[ \sum_{i_1 \geq 0, \ldots, i_m \geq 0, \sum_i = n} j_1! \binom{i_1}{j_1} \cdots j_m! \binom{i_m}{j_m} \binom{n}{i_1, \ldots, i_m} B_{i_1} \cdots B_{i_m} \]

and the form

\[ \sum_{i_1 \geq 0, \ldots, i_m \geq 0, \sum_i = n} \binom{n}{i_1, \ldots, i_m} N_1^{i_1} \cdots N_m^{i_m} B_{i_1}(\alpha_1) \cdots B_{i_m}(\alpha_m), \]

which give rise to various identities involving Bernoulli numbers and polynomials.

2.3.3 Kernel-based discriminant techniques on educational placement (joint with M.H. Lin and Y.C. Chang)

This article considers the problem of educational placement. The data were part of a large scale project for developing educational indicators to monitor and to upgrade Taiwan’s elementary and secondary science education.

A profile vector for each student consists of five science-educational indicators. The students are intended to be placed into three reference groups: advanced, regular and remedial. Kernel-based nonparametric discriminant techniques are introduced. A study of the relative performance of four classification schemes (Fisher’s, kernel-based methods with three bandwidth selection approaches) are investigated in the context of educational placement.
3 Plan of Future Research

3.1 Continuing research.

I will continue to do collaborative research on nonparametric function estimation, Bayesian computation and statistical applications on epidemic studies and educational fields.

- **Wavelet methods.** Adding to our BLUPWAVE sequel, we are now in the process of writing up a complete documentation of the software implementing BLUPWAVE. This part of research is joint with H.S. Lu at Institute of Statistics, National Chiao-Tung University.

- **Statistical applications on epidemic studies.** For the past three years I have been interested and involved in some epidemic studies including longitudinal studies of the progression of myopia among school children and a study of risk assessment of inbreeding among artificial insemination half-sibs. These studies are still ongoing and joint with C.K. Hsiao at Institute of Epidemiology, National Taiwan University.

- **Statistical applications on educational and psychological fields.** I have been working with two colleagues, M.H. Lin and Y.C. Chang, on educational placement problem to monitor and upgrade Taiwan’s elementary and secondary science education. Our joint project is still ongoing.

3.2 Statistical learning theory.

Starting in late 2001, I got interested in the statistical learning theory. Since then a seminar group has been formed with active members from research fields including statistics, computer science and information engineering. The group members have diverse specialties while strong bond to work together. We also seek out cross discipline collaboration opportunity beyond the group. Listed are some specific problems that we are working on.

- **Incremental reduced SVM [19].** Support vector machines have been a popular and important classification algorithm in supervised learning. Via the use of kernel mapping, the SVM allows effective nonlinear classification. There are some major problems that confront large data classification:
  - the size of the mathematical programming problem;
  - the dependence of the nonlinear separating hypersurface on the entire data set, which creates unwieldy storage problems;
  - the overfitting of the separating hypersurface, which causes high variation and less stable fitting.

To overcome these problems Lee and Mangasarian (2000) propose the method of reduced SVM. The key ideas of reduced SVM are as follows. The nonlinear separating
hypersurface is generated by \( K(x, x_i) \) with very few \( x_i \), while the entire data set is used in fitting the hypersurface. This small portion of data, used to generate the hypersurface, is randomly selected. The proposal of reduced SVM by Lee and Mangasarian (2000) works successfully.

Joint with Y.J. Lee at Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, a theoretical study of the reduced SVM is in the process. We show that the random subset selection approach for choosing basis functions \( K(x, x_i) \) is nearly optimal in a certain sense. We study the reduced SVM from the model complexity point of view and propose a further refinement algorithm, named incremental reduced SVM. The incremental reduced SVM is a fast algorithm based on incremented subset selections to carry out SVM for extremely large data set.

- **Equivalence relations between support vector machines, sparse approximation and Bayesian regularization** [20]. Girosi (1998) proved an equivalence relation between support vector regression and a sparse approximation scheme under the assumption that the data have been obtained by sampling without noise. In his paper, Girosi raised an open problem of equivalence for the noisy case. In this article we extend Girosi’s notion of equivalence to being considered in a larger reproducing kernel Hilbert space, which contains as subspaces both the subspace of signals and the subspace generated by the covariance kernel of an error process. Reproducing kernel Hilbert space provides a unified framework for SVM as well as for a wide variety of statistical modelling and function estimation problems (Wahba, 2002). With such a framework, we establish several equivalence relations for variants of support vector machine algorithms for the noisy case, including 1-norm soft margin SVM classification, 2-norm and a modified 2-norm soft margin proximal SVM classification, and support vector regression with \( \epsilon \)-insensitive loss. Certain Bayesian regularization problems and Gauss-Markov prediction in statistical mixed-effects models are found to be linked to SVM algorithms and are also discussed in this article. (Joint with Y.J. Lee.)

- **Bayesian kernel FDA.** This is a joint research with Y.C. Chang. The classical Fisher’s linear discriminant analysis (FDA) is a commonly used and time-honored tool for classification. As the successful development and usage of the kernel method, kernelized versions of FDA have been proposed and studied by some authors. The idea of kernelization is to map the data from the input space into a high-dimensional feature space and then perform the Fisher’s linear discriminant analysis in the feature space. Such a kernel FDA allows flexible nonlinear discriminant function in the input space. For references, see Mika, Rätsch, Weston, Schölkopf and Müller (1999), Baudat and Anouar (2000), Mika, Rätsch and Müller (2001), Van Gestel, Suykens and De Brabanter (2001), Mika, Smola and Schölkopf (2001), Xu, Zhang and Li (2001). The kernel method implicitly defines an embedding of the original data into a feature space, upon which a similarity measure can be defined via the kernel value. Besides
the interpretation of similarity and the convenience of flexible nonlinear discriminant, the kernel method has other interesting features, too. Here we focus on Bayesian aspects of kernelization and interpret the kernelized discriminant hypersurface from a stochastic point of view. The Bayesian approach is a hybrid learning algorithm combining the least squares criterion and function regularization. Using the Bayesian notion, we are able to tune the smoothness, the shape and the disposition of the discriminant hypersurface. Probabilistic outputs are available through the Bayesian posterior analysis.

4 References


