Applications of Convex Optimization in Premium Rating

Dimitri Semenovich

Abstract: In this paper we discuss the application of modern mathematical optimization techniques to some of the common problems in insurance premium rating. The computationally tractable setting of convex optimization [6] is particularly attractive as it encompasses parameter estimation in generalized linear models and offers means to address practical challenges such as variable selection, coefficient smoothing, spatial and hierarchical priors, constraints on relativities and the time evolution of model parameters. Recent advances in modelling systems for convex optimization make these methods not only eminently practical but also in many respects more flexible than what is presently offered by statistical software.

Keywords: Convex optimization, generalized linear models, credibility, graduation, spatial smoothing, dynamic regression, revenue management.

1 Introduction

In this paper we formulate several common models arising in premium rating as convex optimization problems and describe the use of constraints and regularization to address practical issues such as variable selection, coefficient smoothing, hierarchical credibility, parameter evolution and spatial priors in a unified framework. The resulting optimization problems can be solved using efficient algorithms [65] developed for convex programming[1].

Many classical actuarial techniques such as Whittaker graduation and various credibility models can be interpreted as performing regularized or constrained fitting [10, 38, 29]:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathcal{L}(\mathbf{y};\mathbf{w}) + R(\mathbf{w}) \qquad\qquad \begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \mathcal{L}(\mathbf{y};\mathbf{w}) \\ \text{subject to} & R(\mathbf{w}) \leq \epsilon, \end{array} \tag{1}$$

where $\mathcal{L}(\mathbf{y};\mathbf{w})$ is the term penalizing model error relative to the data $\mathbf{y}$ and the regularization term $R(\mathbf{w})$ measures the lack of smoothness or some other desired property of the model $\mathbf{w}$. While not a part of the classical theory, regularization has important implications for the practical use of generalized linear models (GLMs), by now a nearly universal tool in premium rating. Namely, rather than carrying out manual feature design and selection it may often be far more effective to control the degrees of freedom of the model by imposing penalties or constraints on the coefficients. Recent advances in solvers and open source modelling software for convex optimization [21] have made it exceedingly easy to develop such custom models.

The idea of regularization itself has been developed independently many times in many different fields, e.g. the work of A. Tikhonov on operator equations in the 1960s [58]. It is also the principal reason for the remarkable performance of the "support vector machines" family of algorithms which implicitly generate feature spaces of high dimension through the kernel functions [59].

[1] Video lectures for the Stanford course EE364A Convex Optimization, made publicly available via the Stanford Engineering Everywhere initiative, are highly recommended as background for this paper.

In recent years there has been a resurgence of interest in a particular form of regularization known as "ℓ_1-norm" regularization. Its early applications appeared in geophysics in 1970s [55] where its sparsity inducing properties had been employed for signal recovery. Around the same time in the actuarial literature Schuette proposed an ℓ_1-norm formulation of the graduation problem [45] which allowed it to be solved by linear programming; this idea was further developed in econometrics as quantile smoothing splines [35]. It was not until much later that ℓ_1-norm regularization had become a widely adopted technique in signal processing with applications including computing transform coefficients ("basis pursuit") [12] and signal recovery from incomplete measurements ("compressive sensing") [14]. In statistics, the idea of ℓ_1-norm regularization was popularized by the well-known "lasso" procedure [56] for linear regression and its many extensions [57, 68, 67, 37]. While for a long time ℓ_1-norm regularization has been viewed as little more than a useful heuristic in optimization, recent theoretical results (e.g. [9]) have provided surprising guarantees on its performance in certain restricted settings.

What the above models have in common is that they all, together with many others[2] [6], can be formulated as convex optimization problems. Recognizing convexity and its implied properties offers a unifying perspective on a collection of seemingly unrelated ideas from many different fields and dramatically reduces the need to develop special purpose algorithms as many instances can be handled by standard solvers. Moreover, convex problems constitute, perhaps, the widest known class of optimization problems for which exist efficient algorithms guaranteed to find a global solution, making convexity especially desirable when reliable numerical solutions are a requirement.

This paper consists of two main parts. In the first half we outline the basics of mathematical optimization and convex calculus and briefly describe the connection between convex optimization and statistical estimation. In the second part we focus on the application of convex optimization to some of the problems in technical premium rating. We discuss variable selection, curve fitting, spacial clustering and smoothing, additive models, hierarchical credibility, time evolution of model parameters and stochastic optimization.

1.1 Related work

A number of unifying approaches along similar lines have been previously proposed in the actuarial literature but using as the foundation algorithmic developments from statistics, rather than mathematical optimization. These include Bayesian multilevel extensions of generalized linear models [19, 18] and the so called "mixed effect" models [1]. Unlike convex optimization, however, these have arguably less broad scope of applicability (e.g. ability to handle constraints is lacking) and do not take into account computational complexity of the proposed methods.

It should be noted at the same time that stateful sampling algorithms such as variants of geometric Markov chain Monte Carlo [25], used for inference in Bayesian models, bear close resemblance to iterative optimization methods and their hybridization is an active area of research, e.g. [63]. It seems not unreasonable to hope that one day an effective synthesis will be attained.

The most complete treatment of actuarial models from the perspective of convex optimization was developed in a visionary paper of Brockett [7], albeit in the *dual* form[3] and some 15 years before advances in algorithms and computing power have made such schemes truly practical for all insurance applications.

[2] In particular it is worth noting that one of many fields where convex analysis has proven to be the key tool is mathematical economics; see [61] for an accessible introduction from this point of view.

[3] The equivalence between "information theoretic" maximum entropy principle and maximum likelihood estimation for exponential families is discussed in standard references, e.g. [8].

2 Optimization and statistical inference

2.1 Mathematical optimization

A constrained optimization problem has the the following form:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & f_0(\mathbf{x}) \\ \text{subject to} & f_i(\mathbf{x}) \leq b_i,\ i = 1, \ldots, m. \end{array} \tag{2}$$

The vector $\mathbf{x} \in \mathbb{R}^n$ is the optimization variable, the function $f_0 : \mathbb{R}^n \to \mathbb{R}$ is the objective function, the functions $f_i : \mathbb{R}^n \to \mathbb{R},\ i = 1, \ldots, m$ are the inequality constraint functions[4] , and the constants $b_1, \ldots, b_m$ are the limits, or bounds, for the constraints. A vector $\mathbf{x}^*$ is called a *solution* or a *global minimum* of the optimization problem (2) if no other vector satisfying the constraints achieves a smaller objective value, that is for any $\mathbf{z} \in \mathbb{R}^n$ with $f_1(\mathbf{z}) \leq b_1, \ldots, f_m(\mathbf{z}) \leq b_m$ we have $f_0(\mathbf{z}) \geq f_0(\mathbf{x}^*)$.

2.1.1 Convex optimization

While the general problem (2) is computationally completely intractable, we can mitigate this by restricting the class of functions $f_0, \ldots, f_m$. For example, if we take the objective and the constraints to be linear, the optimization problem (2) is called a linear program and can be written as:

$$\begin{array}{ll} \underset{\mathbf{x}}{\text{minimize}} & \mathbf{c}^T\mathbf{x} \\ \text{subject to} & \mathbf{a}_i^T\mathbf{x} \leq b_i,\ i = 1, \ldots, m. \end{array} \tag{3}$$

Despite the seeming simplicity surprisingly many problems in business operations and engineering (e.g. optimal network flow) can be expressed in this form [3]. There are also applications in statistics and econometrics (least absolute deviations, quantile regression [34]). The subject was developed in 1930s and 40s by L. Kantorovich and G. Dantzig. The latter introduced the simplex algorithm for solving linear programs that for many applications remains unsurpassed to this day and can routinely solve problems with millions of variables and constraints.

A less restrictive class of tractable optimization problems is the one in which the objective and constraint functions are convex, namely:

$$f_i(\alpha\mathbf{x} + \beta\mathbf{y}) \leq \alpha f_i(\mathbf{x}) + \beta f_i(\mathbf{y}) \tag{4}$$

for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and all $\alpha, \beta \in \mathbb{R}$ with $\alpha + \beta = 1,\ \alpha \geq 0,\ \beta \geq 0$. Convex functions have the essential property that every local minimum is also a global minimum. A function $g(\mathbf{x})$ is called a concave function if $-g(\mathbf{x})$ is convex. It is easy to check that linear functions are convex (they are also concave) and therefore linear programs are a special case of convex optimization problems.

There are efficient polynomial time algorithms (e.g. so called interior point methods [6]) for global minimization of convex functions subject to convex inequality constraints. Indeed convex problems are effectively the widest class of optimization problems for which such algorithms exist at this time. And while convexity is not an essential property of a successful optimization model, it is worthwhile to be aware of the trade-off between efforts to make a model more realistic and ensuing difficulties with numerical methods. To quote Y. Nesterov [42], one of the key figures in the development of convex programming:

> Every year I meet Ph.D students of different specializations who ask me for advice

[4] Equality constraints are omitted for brevity but are implied, i.e. an equality constraint $f_i(\mathbf{x}) = 0$ can be represented as $f_i(\mathbf{x}) \geq 0$ and $-f_i(\mathbf{x}) \geq 0$.

> on reasonable numerical schemes for their optimization models. And very often they seem to have come too late. In my experience, if an optimization model is created without taking into account the abilities of numerical schemes, the chances that it will be possible to find an acceptable numerical solution are close to zero.

2.1.2 Convex calculus

It is generally labor intensive to check convexity of a function directly from the definition. In most cases it is much easier to see wheteher a given function is built up of known convex functions using transformations that preserve convexity, just as classical calculus permits effective computation without explicitly working with infinite series.

Familiarity with the basics of convex analysis and a few heuristics will permit effective creation of custom models for many applications. The tedious (and ultimately mechanical) task of converting the resulting formulation to one of standard forms understood by solvers can be handled by the modelling system [21].

We describe some convex functions of one and many variables together with operations that maintain convexity in Appendix A. Much more detailed treatments can be found in [44, 6, 2, 42].

2.2 Convex optimization and statistical estimation

Statistical inference can often reduced to solving certain optimization problems. Below we discuss two such principles.

2.2.1 Maximum likelihood and loss minimization:

A familiy of probability density functions on $\mathbb{R}^n$ denoted $p(\mathbf{y};\mathbf{w})$ with parameter vector $\mathbf{w} \in \mathbb{R}^m$ is called a likelihood function when taken as a function of $\mathbf{w}$ only for a fixed $\mathbf{y}$. It is, however, often more convenient to deal with the logarithm of the likelihood function or the log-likelihood, $\log p(\mathbf{y};\mathbf{w})$. The negative log-likelihood is sometimes also called the "loss function":

$$\mathcal{L}(\mathbf{y};\mathbf{w}) = -\log p(\mathbf{y};\mathbf{w}). \tag{5}$$

It is worth noting, however, that not all loss functions are directly motivated by a priori distributional assumptions e.g. quantile loss (27).

A remarkably effective method for estimating the parameter $\mathbf{w}$ given an observation $\mathbf{y}$ consists of maximizing the log-likelihood (equivalently, minimising the loss function) with respect to $\mathbf{w}$:

$$\mathbf{w}^* = \underset{\mathbf{w}}{\operatorname{argmin}}\, \mathcal{L}(\mathbf{y};\mathbf{w}). \tag{6}$$

In many practical applications we have prior information that can be represented in the form of constraints on the admissible values of $\mathbf{w}$. These constraints can be defined explicitly by specifying a set $C \subseteq \mathbb{R}^m$ such that $\mathbf{w} \in C$ or incorporated into the likelihood function by setting $p(\mathbf{y};\mathbf{w}) = 0$ and correspondingly $\mathcal{L}(\mathbf{y};\mathbf{w}) = \infty$ for all $\mathbf{w} \notin C$. When C is given, the constrained maximum likelihood estimation problem can be written as follows:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \mathcal{L}(\mathbf{y};\mathbf{w}) \\ \text{subject to} & \mathbf{w} \in C. \end{array} \tag{7}$$

While computationally intractable in general, maximum likelihood estimation is reduced to a convex optimization problem if the loss function $\mathcal{L}(\mathbf{y};\mathbf{w})$ is convex in $\mathbf{w}$ and C is a convex set.

2.2.2 Bayesian estimation

Maximum likelihood procedure has an analogue in the Bayesian setting known as maximum a posteriori estimation. Here the parameter vector $\mathbf{w}$ and the observation $\mathbf{y}$ are both considered to be random variables with a joint probability density $p(\mathbf{y}, \mathbf{w})$. The density of $\mathbf{w}$ is then given by

$$p(\mathbf{w}) = \int p(\mathbf{y}, \mathbf{w})\, dy_1 \dots dy_n. \tag{8}$$

This is referred to as the prior distribution of $\mathbf{w}$ and represents the information about $\mathbf{w}$ before $\mathbf{y}$ is observed. We can similarly define $p(\mathbf{y})$, the prior distribution of $\mathbf{y}$. The conditional probability density of $\mathbf{y}$ given $\mathbf{w}$ is as follows:

$$p(\mathbf{y}|\mathbf{w}) = \frac{p(\mathbf{y}, \mathbf{w})}{p(\mathbf{w})}. \tag{9}$$

Being a function of $\mathbf{w}$, it is equivalent to the likelihood function in employed in maximum likelihood estimation. The conditional probability density of $\mathbf{w}$ given $\mathbf{y}$ can then be written as:

$$p(\mathbf{w}|\mathbf{y}) = \frac{p(\mathbf{y}|\mathbf{w})p(\mathbf{w})}{p(\mathbf{y})} \tag{10}$$

If we substitute the observed value of $\mathbf{y}$ into $p(\mathbf{w}|\mathbf{y})$ we obtain the posterior density of $\mathbf{w}$, representing the updated information about $\mathbf{w}$. The maximum a posteriori estimate of $\mathbf{w}$ is the one that maximizes the posterior probability ($p(\mathbf{y})$ does not depend on $\mathbf{w}$ and can be omitted):

$$\mathbf{w}^* = \operatorname*{argmax}_{\mathbf{w}} p(\mathbf{w}|\mathbf{y}) = \operatorname*{argmax}_{\mathbf{w}} p(\mathbf{y}|\mathbf{w})p(\mathbf{w}). \tag{11}$$

After taking the logarithm the expression for $\mathbf{w}^*$ can be written as:

$$\mathbf{w}^* = \operatorname*{argmax}_{\mathbf{w}} \ \log\big(p(\mathbf{y}|\mathbf{w})p(\mathbf{w})\big) = \operatorname*{argmin}_{\mathbf{w}} \ -\log\big(p(\mathbf{y}|\mathbf{w})\big) - \log\big(p(\mathbf{w})\big). \tag{12}$$

This is equivalent to minimising a data dependent loss function $-\log(p(\mathbf{y}|\mathbf{w}))$ with the additional regularization term $-\log(p(\mathbf{w}))$.

It is also a basic consequence of Lagrange duality [6, Chapter 5] that under some mild regularity conditions (12) has the same solution $\mathbf{w}^*$ as the following constrained optimization problem for some instance dependent value of ϵ:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & -\log\big(p(\mathbf{y}|\mathbf{w})\big) \\ \text{subject to} & -\log\big(p(\mathbf{w})\big) \le \epsilon. \end{array} \tag{13}$$

For example if the prior density of $\mathbf{w}$ has support over a a set C and is uniform then finding the maximum a posteriori estimate is the same as loss minimization subject to the constraint $\mathbf{w} \in C$.

For any estimation problem with a convex loss function we can add a convex regularization term (corresponding to a prior density on $\mathbf{w}$ that is log-concave) and the resulting optimization problem will be convex.

2.3 Convex loss functions

A large number of statistical problems can be reduced to minimizing convex loss functions, with conditional exponential families being perhaps the key example. Several approaches not based directly on the maximum likelihood principle are also mentioned.

2.3.1 Conditional exponential families

It is a standard result that the log-likelihood of distributions in the exponential families is concave in the natural parameters [8, 62]. Below we discuss convexity properties of conditional exponential families, closely related to generalized linear models.

Consider an exponential family distribution on $\mathcal{Y} \times \mathcal{X}$:

$$\begin{aligned} p(y, x|\mathbf{w}) &= h_0(y, x) \exp\Big(\sum_{k=1}^{m} w_k \phi_k(y, x) - A(\mathbf{w})\Big) \\ &= h_0(y, x) \frac{\exp\Big(\sum_{k=1}^{m} w_k \phi_k(y, x)\Big)}{\exp\big(A(\mathbf{w})\big)}. \end{aligned} \tag{14}$$

In this context the non-negative function h_0 is the base or *carrier* measure, $\mathbf{w} \in \mathbb{R}^k$ are the model parameters, $\phi(y, x) = [\phi_1(y, x), \ldots, \phi_k(y, x)]^T$ is the vector of ***sufficient statistics*** and $A(\mathbf{w})$ is the logarithm of the normalizing constant or the ***log-partition*** function, namely:

$$A(\mathbf{w}) = \log\Big(\int_{(y,x)\in\mathcal{Y}\times\mathcal{X}} \exp\Big(\sum_{k=1}^{m} w_k \phi_k(y, x)\Big) h_0(y, x)\, dx\, dy\Big), \tag{15}$$

with summation replacing the integral for discrete distributions. For reasons such as data and computational limitations, we may instead wish to directly estimate the conditional probability:

$$p(y|x, \mathbf{w}) = h_0(y, x) \exp\Big(\sum_{k=1}^{m} w_k \phi_k(y, x) - A(\mathbf{w}|x)\Big), \tag{16}$$

with conditional log-partition function given by:

$$A(\mathbf{w}|x) = \log\Big(\int_{y\in\mathcal{Y}} \exp\Big(\sum_{k=1}^{m} w_k \phi_k(y, x)\Big) h_0(y, x)\, dy\Big). \tag{17}$$

Note that the sufficient statistics that do not depend on y can effectively be omitted as the choices of associated parameters do not influence the conditional densities. Given a collection of independent samples $(y_i, x_i) \in \mathcal{Y} \times \mathcal{X}$ for $i = 1, \ldots, n$, the joint conditional probability can be written as:

$$\prod_{i=1}^{n} p(y_i|x_i, \mathbf{w}) = \prod_{i=1}^{n} \Big(h_0(y_i, x_i) \exp\Big(\sum_{k=1}^{m} w_k \phi_k(y_i, x_i) - A(\mathbf{w}|x_i)\Big)\Big), \tag{18}$$

giving rise to the following maximum log-likelihood estimation problem to find parameters $\mathbf{w}$:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \sum_{i=1}^{n} \Big(A(\mathbf{w}|x_i) - \sum_{k=1}^{m} w_k \phi_k(y_i, x_i)\Big). \tag{19}$$

The above objective is convex being a sum of linear terms and the convex log-partition functions. The latter are convex in $\mathbf{w}$ by an extension of the soft max rule (see A.3).

In this formulation $\mathcal{Y}$ is not restricted to be equal to $\mathbb{R}$ or to a small set of discrete outcomes but can also represent more complex structured objects, e.g. all possible parts of speech assignments for a particular sentence. This type of models is commonly[5] referred to as a *conditional random field* [49] and is likely to prove quite fruitful in insurance applications.

To recover the (now classical) generalized linear models of Wedderburn and Nelder, consider

[5] At least in computer science literature on "machine learning" (a variant of computational statistics with a strong focus on out of sample performance) and natural language processing.

the case when $\mathcal{X} = \mathbb{R}^k$, the carrier measure h_0 does not depend on x and with a particularly simple choice of sufficient statistics:

$$\phi_k(y, \mathbf{x}) = yx_k. \tag{20}$$

We can then rewrite (16) as a single parameter exponential family with respect to $\mathbf{w}^T\mathbf{x}$ in the following way:

$$\begin{aligned} p(y|\mathbf{x}, \mathbf{w}) &= p(y|\mathbf{w}^T\mathbf{x}) \\ &= h_0(y) \exp\Big(y\mathbf{w}^T\mathbf{x} - B(\mathbf{w}^T\mathbf{x})\Big), \end{aligned} \tag{21}$$

with the log partition function:

$$B(\theta) = \log\Big(\int_{y\in\mathcal{Y}} \exp\left(\theta y\right) h_0(y)\, dy\Big) \tag{22}$$

and giving rise to the following maximum likelihood parameter estimation problem:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathcal{L}(y; X\mathbf{w}) = \sum_{i=1}^{n} \left(B(\mathbf{w}^T\mathbf{x}_i) - y_i\mathbf{w}^T\mathbf{x}_i\right). \tag{23}$$

The usual relation between the natural parameter and the expected value of y obtains:

$$\mathbb{E}(y|\mathbf{x}, \mathbf{w}) = \nabla_\theta B(\mathbf{w}^T\mathbf{x}), \tag{24}$$

with $\nabla_\theta B^{-1}$ being the canonical link function.

There are two remaining incompatibilities between conditional random fields and generalized linear models, however. Firstly GLMs are based on the so called *exponential dispersion* families [31]:

$$p(y|\mathbf{x}, \mathbf{w}, \lambda) = h_0(y, \lambda) \exp\Big(\lambda\big(y\mathbf{w}^T\mathbf{x} - B(\mathbf{w}^T\mathbf{x})\big)\Big), \tag{25}$$

rather than exponential families considered up to this point.

For the fixed dispersion parameter $\sigma^2 = \lambda^{-1}$ this class of models coincides with single parameter exponential families. This is often the case, e.g. when σ^2 represents a known number of observations for the binomial distribution. If σ^2 is not known and is to be estimated, the resulting log-likelihood is not in general jointly concave in w and σ^2, unlike that for conditional random fields. One way to overcome this limitation is to consider the overlapping class of two parameter exponential families, which includes many standard distributions and also provides means to deal with overdispersion [13].

Another difficulty is with regard to the link function - convexity of the negative log-likelihood does not necessarily hold for choices other than the canonical link. Even in this case, however, for all of the models described in this paper local solutions can be obtained using sequential quadratic approximation (variants of which are known as iteratively reweighted least squares and Fisher scoring). It is worth remembering that none of the methods implemented in existing statistical software provide guarantees of global optimality in this situation either.

Convex formulations for parameter estimation in some common GLMs are shown in Appendix B.

2.3.2 Huber loss

Huber loss function [26] is often used to make least squares estimates more robust to outliers, see also [6, Chapter 6] for an illuminating informal discussion. It agrees with the squared ℓ_2-loss (96) for $|u| < M$ and for $|u| \geq M$ the Huber loss function reverts to linear growth which gives lowest attainable sensitivity to outliers while still maintaining convexity.

$$\mathcal{L}(\mathbf{y}; X\mathbf{w}) = \sum_{i=1}^{n} \phi(y_i - \mathbf{w}^T\mathbf{x}_i), \quad \phi(u) = \begin{cases} u^2, & |u| < M \\ M(2|u| - M), & |u| \geq M. \end{cases} \tag{26}$$

2.3.3 Quantile loss

One interpretation of the least squares procedure (96) is that it estimates the conditional mean of y_i given the data vector $\mathbf{x}_i$. Regression with the asymmetric quantile [34] loss function ρ_τ on the other hand results in estimates approximating the conditional τ-th quantile of the response variables y_i:

$$\mathcal{L}(\mathbf{y}; X\mathbf{w}) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{w}^T\mathbf{x}_i), \quad \rho_\tau(u) = \begin{cases} \tau u, & u > 0 \\ -(1-\tau)u, & u \leq 0. \end{cases} \tag{27}$$

When τ is equal to 0.5 and corresponds to the median, quantile regression is equivalent to the method of least absolute deviations which estimates $\mathbf{w}$ by seeking to minimize $\frac{1}{2}\|\mathbf{y} - X\mathbf{w}\|_1$. We should note that quantile regression appears quite attractive for insurance applications [36] as it can provide a non-parametric estimate of the full conditional distribution of the dependent variable and can deal with such issues as concetration of probability mass at a certain point. In addition to directly modelling claim costs per policy, it may also be applicable in situations when mean of the dependent variable is not easily interpreted, e.g. when fitting a model to a sample of competitor rates.

3 Applications in premium rating

3.1 Variable selection

One compelling application of constrained parameter estimation is variable selection. Consider a regression type problem with an arbitrary convex loss function $\mathcal{L}(\mathbf{y}; X\mathbf{w})$ where y is the vector of response variables, X is the design matrix and w a vector of parameters to be estimated. This can be accomplished by restricting the ℓ_1-norm of the coefficient vector (assuming that the explanatory variables are standardized with mean 0):

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \mathcal{L}(\mathbf{y}; X\mathbf{w}) \\ \text{subject to} & \|\mathbf{w}\|_1 \leq \epsilon. \end{array} \tag{28}$$

This generally results in a sparse estimate $\mathbf{w}^*$ with the number of non zero entries controlled by the magnitude of ϵ. It is possible to motivate this formulation as a convex approximation (or *relaxation*) of the computationally intractable best subset selection problem [6]. Using an ℓ_1-norm penalty to obtain a sparse solution has been a well known heuristic in optimization and its applications going back at least to the 1970s [55, 45]. It was popularized in the statistics literature as the "lasso" by Tibshirani [56]. Under some mild conditions the above problem can be equivalently formulated as:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathcal{L}(\mathbf{y}; X\mathbf{w}) + \lambda\|\mathbf{w}\|_1 \tag{29}$$

where $\lambda \geq 0$ corresponds to the optimal Lagrange multiplier associated with the inequality constraint in (28). In this form the "lasso" procedure has a Bayesian interpretation as a maximum a posteriori estimate with a Laplacian prior on $\mathbf{w}$ with mean zero and variance $\frac{2}{\lambda^2}$. The value of λ controls the sparsity of $\mathbf{w}*$ and can be chosen via cross-validation.

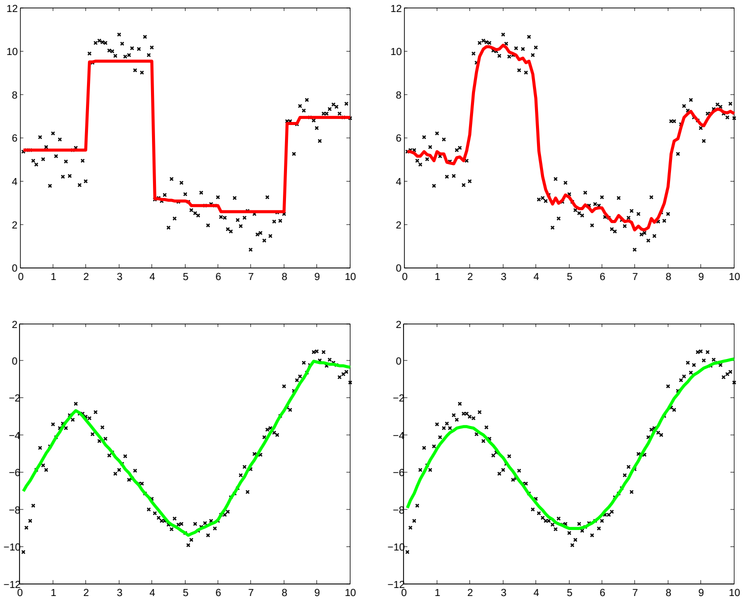


Figure 1: *Left:* Solutions of optimization problem (39) for different choices of D - "fused lasso", $D^{(1,n)}$ (*top*) and "ℓ_1-trend filtering", $D^{(2,n)}$ (*bottom*). *Right:* Solutions of the Whittaker graduation problem (32) with the same choices of D. We set λ to give the same squared error $\|y - w\|_2^2$ as the corresponding solutions on the left.

In the presence of correlated covariates "lasso" will tend to select only one of them. To alleviate this problem we can trade off between penalising ℓ_1 and squared Euclidean norms of $\mathbf{w}$:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathcal{L}(\mathbf{y}; X\mathbf{w}) + \lambda\left(\alpha\|\mathbf{w}\|_1 + (1-\alpha)\|\mathbf{w}\|_2^2\right), \tag{30}$$

where $0 \leq \alpha \leq 1$. This method is known as "elastic net" [68] in the statistics literature. The "elastic net" performs variable selection while at the same time pushing together coefficient values of correlated variables. Indeed, for $\alpha = 0$ it is equivalent to ridge regression, a classical procedure for dealing with collinearity:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathcal{L}(\mathbf{y}; X\mathbf{w}) + \lambda\|\mathbf{w}\|_2^2, \tag{31}$$

see [40] for connections with credibility.

3.2 Graduation or curve fitting

3.2.1 One dimensional data

Before we discuss further extensions to generalized linear models we motivate our approach by examining the classical setting of non-parametric graduation. The goal of the graduation procedure is to smooth a sequence of observations $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$ which are usually indexed by time or age. As pointed out in [46], as early as 1899 Bohlmann had proposed [4] to perform graduation by solving the following convex optimization problem:

$$\underset{\mathbf{w}}{\text{minimize}} \; \|\mathbf{y} - \mathbf{w}\|_2^2 + \lambda\|D\mathbf{w}\|_2^2, \tag{32}$$

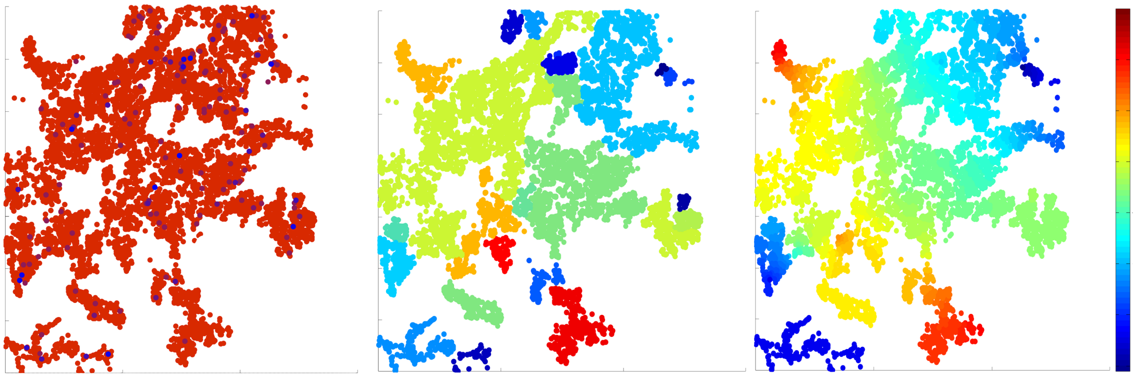


Figure 2: *Left:* Original residuals (motor portfolio). *Center:* Solution to (39) using the incidence matrix of the 10 nearest neighbors graph for regularization. *Right:* Solution using the graph Laplacian.

with $\lambda \geq 0$ and $D = D^{(1,n)}$ being the $(n-1) \times n$ first order finite differences matrix:

$$D^{(1,n)} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \cdots & & \\ & & -1 & 1 \end{bmatrix}. \tag{33}$$

Heuristically, the "fidelity" term $\|\mathbf{y} - \mathbf{w}\|_2^2$ encourages the solution $\mathbf{w}$ to be close to the original data $\mathbf{y}$ and the smoothness term $\|D^{(1,n)}\mathbf{w}\|_2^2$ penalizes non-zero entries of $D^{(1,n)}\mathbf{w}$, first order finite differences (or the discretized first derivative) of $\mathbf{w}$. The value of the parameter λ determines the relative importance of the smoothness term. By Lagrange duality we can obtain the same solution as for (32) with any value of $\lambda \geq 0$ by solving either of the following constrained optimization problems for some values of ϵ_1 and ϵ_2:

$$\begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \|\mathbf{y} - \mathbf{w}\|_2^2 \\ \text{subject to} & \|D\mathbf{w}\|_2^2 \leq \epsilon_1 \end{array} \qquad \begin{array}{ll} \underset{\mathbf{w}}{\text{minimize}} & \|D\mathbf{w}\|_2^2 \\ \text{subject to} & \|\mathbf{y} - \mathbf{w}\|_2^2 \leq \epsilon_2, \end{array} \tag{34}$$

i.e. the objective and the constraint can be freely interchanged.

Bohlmann's procedure (32) can be extended to penalize k-th order finite differences. In this case the k-th order finite differences matrix $D^{(k,n)} \in \mathbb{R}^{(n-k)\times n}$ can be defined recursively:

$$D^{(k,n)} = D^{(1,n-k)}D^{(k-1,n)},\ k = 2, 3, \ldots \tag{35}$$

The second order finite differences matrix would then, for example, is as follows:

$$D^{(2,n)} = \begin{bmatrix} -1 & 2 & -1 & & & \\ & -1 & 2 & -1 & & \\ & & \cdots & & & \\ & & -1 & 2 & -1 \end{bmatrix}. \tag{36}$$

Whittaker [64] had described the underlying probabilistic model and an approximate solution method for the case of third order differences and weighted fidelity term. A Bayesian interpretation of Whittaker graduation is also given by Taylor [51]. Among many extensions to Whittaker graduation most relevant to the present discussion are the work of Schuette [45] and Chan et al. [11, 10]. In his remarkable paper, Schuette proposed the formulation using ℓ_1-norm penalties (weights omitted to simplify presentation):

$$\underset{\mathbf{w}}{\text{minimize}} \ \|\mathbf{y} - \mathbf{w}\|_1 + \lambda \|D^{(k,n)}\mathbf{w}\|_1, \tag{37}$$

After applying a standard transformation this optimization problem can be reformulated as a linear program. In the discussion following [45], S. Klugman had pointed out that the method attempts to make most of the entries of $D^{(k,n)}\mathbf{w}$ (or k-th order differences of $\mathbf{w}$) zero but several of them could be large. This "sparsity inducing" property of ℓ_1-norm penalty motivates its use in the "lasso" variable selection procedure 28.

Chan et al. [10] show that for $p, q \in \{1, 2, \infty\}$ the mixed ℓ_p and ℓ_q norm graduation problem:

$$\underset{\mathbf{w}}{\text{minimize}} \ \|\mathbf{y} - \mathbf{w}\|_p + \lambda \|D^{(k,n)}\mathbf{w}\|_q, \tag{38}$$

can be formulated as a linear program whenever $p, q \in \{1, \infty\}$ and as a quadratic program whenever either p or q is 2.

Smoothing techniques equivalent to Whittaker graduation are known under different names in many fields e.g. "Hodrick-Prescott filter" in economics [24]. More recently, a variant of (38):

$$\underset{\mathbf{w}}{\text{minimize}} \ \|\mathbf{y} - \mathbf{w}\|_2^2 + \lambda \|D^{(1,n)}\mathbf{w}\|_1, \tag{39}$$

with $p = 2$ (or equivalently squared ℓ_2-norm) and $q = 1$ has been popularized in applied statistics literature as "fused lasso" [57]. In signal processing the same formulation is called "total variation denoising". This procedure usually gives a piecewise constant solutions $\mathbf{w}^*$ i.e. discrete first derivative $D^{(1,n)}\mathbf{w}^*$ has mostly zero entries (see top left section of Fig. 1). Similarly, using second order differences $D^{(2,n)}$ often results in a piecewise linear $\mathbf{w}^*$ (see bottom left panel of Fig. 1) and has been described as "ℓ_1 trend filtering" [33] and "quantile splines" [35], the latter replacing the quadratic term with quantile loss. These are effective approaches to change point detection and are considerably simpler than many methods proposed to date.

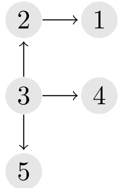
3.2.2 Multidimensional data, spatial smoothing and clustering

We can also apply ℓ_q-norm penalized formulation (38) in situations when the samples $\mathbf{y}$ are over a regular grid or indeed an arbitrary graph (e.g. a k-nearest neighbors graph for objects embedded in a metric space).

To obtain a piecewise constant solution we can use $q = 1$ and the graph incidence matrix for regularization (instead of the first order finite differences matrix $D^{(1,n)}$ in one dimensional case). The incidence matrix A for a graph with n nodes and m edges is a $m \times n$ matrix, with each row representing an edge and composed of a 1 and a -1 in the columns corresponding to the two connected nodes and zeroes elsewhere. See Figure 3 for an example.

To get an equivalent of graduation with second order differences over a regular grid we can consider horizontal and vertical second order differences. As in the one dimensional case, we minimize a weighted sum of the fitting error and,for $q = 1$, a penalty on absolute value of slope changes in the horizontal and vertical directions. The resulting solution tends to be affine over connected regions. The boundaries between regions can be interpreted as curves along which the gradient of the underlying function changes quickly. The approach can not be extended directly to arbitrary graphs. Instead we can use the suitably normalized graph Laplacian (which can be interpreted as a discretization of the Laplace operator) defined as:

$$L = A^T A. \tag{40}$$



$$A = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix} \qquad L = A^T A = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 \\ -1 & 2 & -1 & 0 & 0 \\ 0 & -1 & 3 & -1 & -1 \\ 0 & 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 0 & 1 \end{bmatrix}$$

Figure 3: A simple directed graph (*left*), its incidence matrix A (*center*) and Laplacian L (*right*).

Figure 2 shows the results of ℓ_1-norm spatial smoothing applied to geocoded residuals using the incidence matrix of 10 nearest neighbors graph and its Laplacian. The former provides a piecewise constant solution over connected regions, where the regions with the constant fitted value can be interpreted as clusters, and the latter a piecewise affine solution. In the actuarial literature the use of Whittaker graduation to perform spatial smoothing with irregular regions has been proposed by Taylor [53].

It is worth pointing out that the formulation (32) is in fact quite general, given a free choice of matrix D, as it can viewed as a reparametrization of regularized regression. So for the squared ℓ_2 norm penalty the following problems are equivalent (taking $\mathbf{a} = X\mathbf{w}$ and provided X has full row rank):

$$\underset{\mathbf{w}}{\text{minimize}} \quad \|\mathbf{y} - X\mathbf{w}\|_2^2 + \lambda\|\mathbf{w}\|_2^2, \qquad \underset{\mathbf{a}}{\text{minimize}} \quad \|\mathbf{y} - \mathbf{a}\|_2^2 + \lambda\|\mathbf{a}\|_K^2, \tag{41}$$

where $\|\mathbf{a}\|_K^2 = \mathbf{a}^T K\mathbf{a} = \|D\mathbf{a}\|_2^2$ for $K = (XX^T)^{-1} = D^T D$. The second form is known as "Gaussian process regression" [43], or "kriging"[48] in geostatistics literature.

3.3 Additive models

Additive models were first introduced, perhaps, by Ezekiel in 1920s [15] and extended by Hastie some 60 years later [22, 23]. Generalized linear models form predictions based on a linear function of the features:

$$g(\mathbb{E}y) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_m x_m, \tag{42}$$

where $x_i \in \mathbb{R},\ i = 1, \ldots, m$ are explanatory variables, g is the link function and μ is the expected value of the dependent variable y. Additive models replace the linear combination with a sum of arbitrary functions of explanatory variables:

$$g(\mathbb{E}y) = w_0 + f_1(x_1) + f_2(x_2) + \cdots + f_m(x_m). \tag{43}$$

The richer class of models can lead to overfitting without suitable regularization. The latter is usually achieved by requiring the functions f_i to be (piecewise) smooth.

A similar approach is often followed in the practical applications of GLMs. Continuous variables are discretized into a number of bins. For ordered categorical variables, "natural" levels can be used or if the number of levels is deemed too large, binning is applied to reduce the number of distinct levels. If the total number of bins is not controlled this approach can lead to overfitting and poor predictive performance so it is a standard practice to examine the initial fit and then either manually reduce the number of categories or select a suitable collection of basis functions (e.g. linear or cubic splines) [52].

We observe that it is possible to largely automate this process by casting it into the convex optimization framework and taking advantage of regularization (see Figure 4 for an example). Let $X \in \mathbb{R}^{n\times m}$ be the design matrix for the original problem, then we follow the standard procedure and transform the data by binning each feature into k intervals of equal length (we assume the same number of intervals for every feature for simplicity). This gives a new design matrix $X' \in \mathbb{R}^{n\times km}$

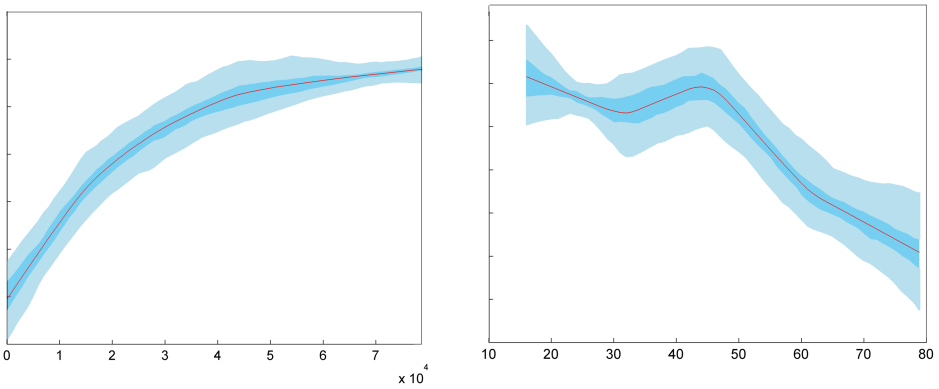


Figure 4: *Left:* The additive effect of sum insured on claims frequency in a motor portfolio. The outer band shows the 90% confidence interval calculated via the bootstrap. *Right:* The additive effect of policyholder age. The anomaly around age 45 could be, for example, due to teenage children driving the family car (also see Figure 5).

and a new parameter vector $\mathbf{w} \in \mathbb{R}^{km}$, with $\mathbf{w}_i \in \mathbb{R}^k$:

$$X' = \begin{bmatrix} X'_1 & , \dots, & X'_m \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \cdots \\ \mathbf{w}_m \end{bmatrix} \tag{44}$$

where X'_p is defined as follows (with $[X]_{i,j}$ denoting x_{ij}):

$$[X'_p]_{i,j} = \begin{cases} 1, & [X]_{i,p} \text{ falls into the } j\text{-th bin} \\ 0, & \text{otherwise} \end{cases} \tag{45}$$

Each observation is now transformed into a sparse vector of dimension km with m non-zero terms. While we lose some information about the features, we can now model non-linear effects in each coordinate. We can write down the parameter estimation problem as:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y}; X'\mathbf{w}) \\ \text{subject to} \quad & \|D\mathbf{w}\|_q \leq \epsilon \end{aligned} \tag{46}$$

or in the equivalent Lagrangian form for some problem specific value of λ:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda \|D\mathbf{w}\|_q \tag{47}$$

where $\mathcal{L}(\mathbf{y}; X'\mathbf{w})$ is an exponential family negative log-likelihood or some other convex loss function and D is a block matrix which evaluates discretized derivatives of the coefficients for each binned explanatory variable:

$$D = \begin{bmatrix} D_1 & & & \\ & D_2 & & \\ & & \cdots & \\ & & & D_m \end{bmatrix}. \tag{48}$$

As shown in section 3.2, we can choose appropriate D_i depending on the structure of the problem and our objectives. Finite difference matrices of up to third order are likely to be sufficient for most applications. The parameter λ can then be chosen by cross-validation. Sometimes it may also

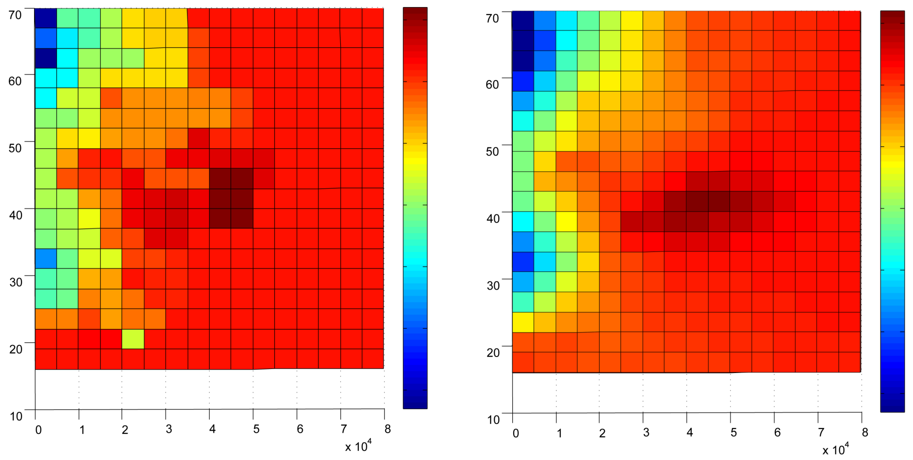


Figure 5: *Left:* Sum insured vs. age interaction for claim frequency in a motor portfolio (same data as Figure 4) using the incidence matrix penalty ("fused lasso"). *Right:* Same interaction but penalising the second order discrete derivative in both coordinates.

be appropriate to introduce independent smoothing parameters (but will make their estimation more difficult):

$$\underset{\mathbf{w}}{\text{minimize}}\ \ \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda_1\|D_1\mathbf{w}_1\|_q + \ \dots\ + \lambda_m\|D_m\mathbf{w}_m\|_q. \tag{49}$$

Notably we can still perform variable selection analogous to the standard ℓ_1-norm regularisation (28) by adding a "group lasso" penalty [67]:

$$\underset{\mathbf{w}}{\text{minimize}}\ \ \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda\|D\mathbf{w}\|_q + \mu\sum_{i=1}^{m}\|\mathbf{w}_i\|_2. \tag{50}$$

Here the term $\sum_{i=1}^{m}\|\mathbf{w_i}\|_2$ encourages the individual components of w to go uniformly to zero as the parameter μ is increased.

3.3.1 Variable interactions

In a Gaussian additive model with identity link function the effect of all the explanatory variables is a sum of their individual effects. Individual effects show how the expected response varies as any single explanatory variable is changed with the others held constant at arbitrary values. For example in order to maximize the expected response we only need to separately maximize each of the component functions of the additive model.

In general there are no guarantees that an additive model will provide a satisfactory fit in any given situation. Non-additivity means that, as one explanatory variable is varied, the effect on the expected response depends on the fixed values of the other variables. Below is an example of how we would change equation (43) if the model is non-additive in variables x_1 and x_2:

$$g(\mu) = w_0 + h(x_1, x_2) + f_3(x_3) + \cdots + f_m(x_m). \tag{51}$$

We can model non-additivity by including the corresponding interactions. Using the notation from equation 45 we can define the interaction $X'_{(1,2)}$ as:

$$[X'_{(1,2)}]_{i,*} = [X'_1]_{i,*} \otimes [X'_2]_{i,*}, \tag{52}$$

where e.g. $[X'_1]_{i,*}$ denotes the i-th row of X'_1 and $\otimes$ is the Kronecker product.

In the resulting model, we can penalize first, second or higher order derivatives in x_1 and x_2 coordinates (see Figure 5), cf. smoothing over a regular grid in section (3.2). Another possibility is to work with the graph Laplacian, which in this case has a particularly simple form:

$$L_{(1,2)} = I \otimes D' + D' \otimes I, \tag{53}$$

with D' a $k \times k$ variant of the second differences matrix $D^{(2,k)}$ with zero padding.

3.3.2 Spatial smoothing and clustering

As geographic regions are not usually arranged in a regular grid the approach in the previous section is not directly applicable for spatial smoothing. Instead, we can introduce indicator variables for each geographic grouping present in the data (postcode, census zone or even unique coordinates) and use a suitably constructed graph representing distance and adjacency information for regularization. Depending on whether we use the graph incidence matrix or the Laplacian, for $q = 1$ we can obtain either a piecewise constant or a smoothly varying surface (see Figure 2) jointly with the other model parameters:

$$\underset{\mathbf{w}}{\text{minimize}} \;\; \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda \left\| D\mathbf{w} \right\|_1, \tag{54}$$

where D is a block diagonal matrix (cf. equation 48) and one of the blocks is either the aforementioned graph incidence matrix or the graph Laplacian. As before, the value of λ can be chosen by cross-validataion.

3.4 Kalman filter and dynamic models

Kalman filter [32] and related ideas have played a central role in the development of state space methods in engineering control through out 1960s (culminating in the linear quadratic Gaussian theory). Remarkably, the first practical application of the Kalman filter was to improve the accuracy of navigation for the Apollo program, quickly followed by adoption for a wide range of aerospace problems. In these applications the goal is typically to track the "state" of a missile or a spacecraft obeying Newtonian dynamics. The state vector would contain the current position together with velocity and acceleration vectors and the goal would be to repeatedly re-estimate the state using measurements coming in from a range of sensors, such as inertial, optical, ground based radar etc.

Kalman filter also has quite a long history in the actuarial literature both as a generalization of the classical linear credibility models [39, 30] and applied to optimal updating of claim reserves [29, 54]. Below we formulate the Kalman filter as an optimization problem and describe some intuitive extensions that make the technique directly applicable to tasks such as ongoing monitoring of conversion rates, claim frequencies or other aspects of portfolio performance.

3.4.1 Kalman filter as an optimization problem

Consider the standard least squares problem (see also Appendix B.1):

$$\underset{\mathbf{w}}{\text{minimize}} \;\; \|\mathbf{y} - X\mathbf{w}\|_2^2, \tag{55}$$

and partition the design matrix X and the response vector $\mathbf{y}$ into m row blocks, corresponding to time periods:

$$X = \begin{bmatrix} X_1 \\ \cdots \\ X_m \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \cdots \\ \mathbf{y}_m \end{bmatrix}. \tag{56}$$

We can then equivalently transform (55) by introducing m copies of the parameter vector and some linear equality constraints:

$$\begin{aligned} &\underset{\mathbf{w}_1,\ldots,\mathbf{w}_m}{\text{minimize}} && \sum_{t=1}^{m} \|\mathbf{y}_t - X_t\mathbf{w}_t\|_2^2 \\ &\text{subject to} && \mathbf{w}_{t+1} - \mathbf{w}_{t+1} = \mathbf{0}, \;\; t = 1, \ldots, m-1. \end{aligned} \tag{57}$$

The above model can also be written in the state space form with the identity state transition matrix, no state transition noise and i.i.d. Gaussian observation noise[6], where $\mathbf{w}_t$ is the unobserved state vector and $\mathbf{y}_t$ are the observations associated with time dependent observation matrices X_t:

$$\begin{aligned} \mathbf{w}_{t+1} = \mathbf{w}_t, \qquad & \mathbf{y}_t = X_t\mathbf{w}_t + \boldsymbol{\epsilon}_t, \\ & \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, I). \end{aligned} \tag{58}$$

By introducing i.i.d. Gaussian state transition noise:

$$\begin{aligned} \mathbf{w}_{t+1} = \mathbf{w}_t + \boldsymbol{\nu}_t, \qquad & \mathbf{y}_t = X_t\mathbf{w}_t + \boldsymbol{\epsilon}_t, \\ \boldsymbol{\nu}_t \sim \mathcal{N}(0, I), \qquad & \boldsymbol{\epsilon}_t \sim \mathcal{N}(0, I) \end{aligned} \tag{59}$$

we effectively relax the equality constraints in (57) replacing them with a squared ℓ_2-norm penalty term. It is then possible to perform estimation by solving the following convex optimisation problem:

$$\underset{\mathbf{w}_1,\ldots,\mathbf{w}_m}{\text{minimize}} \;\; \sum_{t=1}^{m} \|\mathbf{y}_t - X_t\mathbf{w}_t\|_2^2 + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2, \tag{60}$$

which amounts to substituting every independent variable in the model by its interaction with the time index variable t and regularizing the differences in the corresponding parameters. The estimation problem can be either solved directly or transformed back to the standard least squares form:

$$\underset{\mathbf{w}}{\text{minimize}} \;\; \|\mathbf{y}' - X'\mathbf{w}\|_2^2, \tag{61}$$

where the design matrix X' and the response vector $\mathbf{y}'$ are redefined as follows:

$$X' = \begin{bmatrix} X_1 & & & \\ -I & I & & \\ & \cdots & & \\ & & -I & I \\ & & & X_m \end{bmatrix}, \quad \mathbf{y}' = \begin{bmatrix} \mathbf{y}_1 \\ 0 \\ \cdots \\ 0 \\ \mathbf{y}_m \end{bmatrix}, \quad \mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \cdots \\ \mathbf{w}_m \end{bmatrix}. \tag{62}$$

The formulation in (60) simultaneously performs both "filtering" and "smoothing" conditional on all the observations up to time m. If new information becomes available the augmented optimization problem should be solved again to obtain new estimates of the entire history of state transitions $\mathbf{w}^* = [\mathbf{w}_1^*, \ldots, \mathbf{w}_m^*, \mathbf{w}_{m+1}^*]^T$. Standard Kalman filter given information up to time $m+1$, on the other hand, only updates the estimate of the current state $\mathbf{w}_{m+1}^*$ and requires a backward

[6] We can in fact avoid the Gaussianity assumption by positing a quadratic loss function instead. The estimators obtained are the same in both cases.

"smoothing" pass to update estimates of past states $\mathbf{w}_1, \ldots, \mathbf{w}_m$.

Indeed, the Kalman filter followed by a "smoothing" step can be viewed as a computationally efficient recursive procedure for solving the normal equations of the least squares problem (61) which exploits block tri-diagonal structure of the matrix X'. With advances in numerical linear algebra routines for sparse matrices and increasing computer speeds very large problems of this kind can be solved directly.

So far we focused on the special case with the identity state transition matrix and i.i.d noise, however state estimation in the the general linear Gaussian state space model:

$$\begin{aligned} \mathbf{w}_{t+1} &= F\mathbf{w}_t + \boldsymbol{\nu}_t, & \mathbf{y}_t &= X_t\mathbf{w}_t + \boldsymbol{\epsilon}_t, \\ \boldsymbol{\nu}_t &\sim \mathcal{N}(0, \Sigma_\nu), & \boldsymbol{\epsilon}_t &\sim \mathcal{N}(0, \Sigma_\epsilon) \end{aligned} \tag{63}$$

can also be expressed as a convex optimisation problem. Denoting $\|\mathbf{a}\|_P = (\mathbf{a}^T P\mathbf{a})^{\frac{1}{2}}$, P-quadratic norm for a positive definite matrix P, it is:

$$\underset{\mathbf{w}_1,\ldots,\mathbf{w}_m}{\text{minimize}} \quad \sum_{t=1}^{m} \|\mathbf{y}_t - X_t\mathbf{w}_t\|_{\Sigma_\epsilon^{-1}}^2 + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - F\mathbf{w}_t\|_{\Sigma_\nu^{-1}}^2. \tag{64}$$

Notably we can recover both Whittaker graduation and Jones-Gerber "evolutionary" credibility [28] as state space models by choosing a one dimensional state vector w_t with the observation matrix X_t a constant vector $\mathbf{1}$ [30, 60]:

$$\underset{w_1,\ldots,w_m}{\text{minimize}} \quad \sum_{t=1}^{m} \|\mathbf{y}_t - \mathbf{1}w_t\|_2^2 + \sum_{t=1}^{m-1} \|w_{t+1} - w_t\|_2^2. \tag{65}$$

When there is only a single observation per time step, this is identical to Whittaker graduation with first order differences as in (32), while the Gerber-Jones model allows multiple observations per time period.

3.4.2 Some extensions to the dynamic models:

We can use any convex loss for the observations, such as quantile or logistic, and still end up with a convex optimization problem:

$$\underset{\mathbf{w}_1,\ldots,\mathbf{w}_m}{\text{minimize}} \quad \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t\mathbf{w}_t) + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2. \tag{66}$$

In the context of dynamic generalized linear models this corresponds to posterior mode estimation as proposed by Fahrmeier [17, 16]. Non-Gaussian state noise is another possibility. It may be appropriate to apply ℓ_1-norm penalty to state changes provided most of the time parameters stay constant with occasional large jumps (cf. Figure 1):

$$\underset{\mathbf{w}_1,\ldots,\mathbf{w}_m}{\text{minimize}} \quad \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t\mathbf{w}_t) + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_1. \tag{67}$$

Another possibility is a combination of norms, e.g. an approach combining squared ℓ_2-norm and ℓ_1-norm penalties will attempt to decompose the state trajectory into a smooth and a piecewise constant component:

$$\underset{\mathbf{w},\mathbf{c}}{\text{minimize}} \quad \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t(\mathbf{w}_t + \mathbf{c}_t)) + \lambda \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_1 + \mu \sum_{t=1}^{m-1} \|\mathbf{c}_{t+1} - \mathbf{c}_t\|_2^2. \tag{68}$$

We can also allow linear trends in the parameters (this formulation can be reduced to the standard state space model by expanding the state vector):

$$\underset{\mathbf{w}_1,\ldots,\mathbf{w}_m}{\text{minimize}} \quad \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t \mathbf{w}_t) + \sum_{t=1}^{m-2} \|\mathbf{w}_{t+2} - 2\mathbf{w}_{t+1} + \mathbf{w}_t\|_1. \tag{69}$$

Finally it is possible to add arbitrary convex inequality and linear equality constraints (see section 3.5.2), so e.g. seasonality adjustments can be handled by introducing new variables $\mathbf{c}_1, \ldots, \mathbf{c}_m$ and equality constraints:

$$\begin{array}{ll} \underset{\mathbf{w},\mathbf{c}}{\text{minimize}} & \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t(\mathbf{w}_t + \mathbf{c}_t)) + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\ \text{subject to} & \mathbf{c}_t = \mathbf{c}_{t+k}, \ \ t = 1, \ldots, m-k \\ & \sum_{t=1}^{m} \mathbf{c}_t = 0. \end{array} \tag{70}$$

Formulating state space models as regularized regression can make them considerably more intuitive for those lucky not to have a background in control theory.

3.5 Other applications

3.5.1 Hierarchical credibility

We can describe hierarchical credibility [27, 55] models (closely related to both "random effects" from statistics literature [41, 1] and linear filtering [30, 39]) in the optimization framework provided variances are known. This can be achieved by introducing additional variables to the optimization problem. Consider a simple setup with two risks I_1 and I_2 with observations $y_i, \ i \in I_j$, for $j \in \{1, 2\}$ with unit variance and the following group mean:

$$\mathbb{E}y_i = w_j, \quad \forall i \in I_j, \tag{71}$$

where w_j are themselves random variables with group mean w and known variance:

$$\mathbb{E}w_j = w, \ \ \mathrm{Var}(w_j) = \sigma^2, \quad j = 1, 2. \tag{72}$$

We can then obtain the linear credibility estimator as a solution of the following optimization problem:

$$\underset{w_1,w_2,w}{\text{minimize}} \quad \sum_{j=1}^{2} \sum_{i \in I_j} (y_i - w_j)^2 + \frac{1}{\sigma^2} \sum_{j=1}^{2} (w_j - w)^2. \tag{73}$$

While the notation above is rather cumbersome, complex models are easy to implement in practice. This can be achieved by augmenting the data with indicator variables for the lowest level of the hierarchy and regularizing by the squared Euclidean norm of the product of the (tree) graph incidence matrix representing the hierarchical structure and the parameter vector.

3.5.2 Constraints on relativities

Treating maximum likelihood estimation of GLM parameters as a convex optimization problem allows us to introduce arbitrary convex constraints on rate relativities in addition to various smoothness penalties described earlier. One such constraint in the classical GLM theory is the "offset", used e.g. to allow for exposure in Poisson models. It amounts to setting the coefficient associated with the offset term x_i equal to one:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y}; X\mathbf{w}) \\ \text{subject to} \quad & w_i = 1. \end{aligned} \tag{74}$$

Given the range of practical insurance applications of this simple device [66], the overall approach seems promising.

Consider for example an additive model with binned variables and the associated regularization term (47). Possible constraints would include bounds on the absolute magnitude of the effect $\mathbf{w}_i$ associated with the i-th risk factor or its rate of growth (first differences):

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda\|D\mathbf{w}\|_1 & \qquad \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda\|D\mathbf{w}\|_1 \\ \text{subject to} \quad & \mathbf{w}_i \geq -\epsilon_1\mathbf{1} & \qquad \text{subject to} \quad & D^{(1,k)}\mathbf{w}_i \geq -\epsilon_2\mathbf{1} \\ & \mathbf{w}_i \leq \epsilon_1\mathbf{1} & & D^{(1,k)}\mathbf{w}_i \leq \epsilon_2\mathbf{1}. \end{aligned} \tag{75}$$

In the above $\geq$ denotes entrywise vector inequality and the constraints can alternatively be written as:

$$|w_{ij}| \leq \epsilon_1, \;\; i = 1, \ldots, k \qquad\qquad |w_{i,j+1} - w_{ij}| \leq \epsilon_2, \;\; i = 1, \ldots, k-1. \tag{76}$$

We can also directly control the shape of the additive effect in a particular variable, requiring it, for example, to be non-decreasing or convex (these two might be appropriate for sum insured and driver age respectively):

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda\|D\mathbf{w}\|_1 & \qquad \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y}; X'\mathbf{w}) + \lambda\|D\mathbf{w}\|_1 \\ \text{subject to} \quad & D^{(1,k)}\mathbf{w}_i \geq \mathbf{0} & \qquad \text{subject to} \quad & D^{(2,k)}\mathbf{w}_i \geq \mathbf{0} \end{aligned} \tag{77}$$

Here monotonicity constraints[7] can be written without matrix notation as:

$$w_{i1} \leq w_{i2} \leq \; \ldots \; \leq w_{ik} \tag{78}$$

and similarly for convexity we obtain:

$$w_{ij} - 2w_{i,j+1} + w_{i,j+2} \quad j = 1, \ldots, k-2. \tag{79}$$

Finally consider a problem where there are two known risks ξ_1 and ξ_2 for which we want the percentage difference in premium to be within a certain range ϵ (either due to market or regulatory considerations). Assuming the model uses the log link, this yields:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimize}} \quad & \mathcal{L}(\mathbf{y}; X\mathbf{w}) \\ \text{subject to} \quad & |\mathbf{w}^T\xi_1 - \mathbf{w}^T\xi_2| \leq \log(1+\epsilon). \end{aligned} \tag{80}$$

3.5.3 Demand based pricing

In recent years optimization has most often been mentioned in insurance context in relation to various aspects of "demand driven" pricing which amounts to varying margins in order to maximize underwriting profit. Below we present a very simple model for pricing renewals which admits

[7] Remarkably, there are entire monographs devoted to this topic under the name of *isotonic regression.*

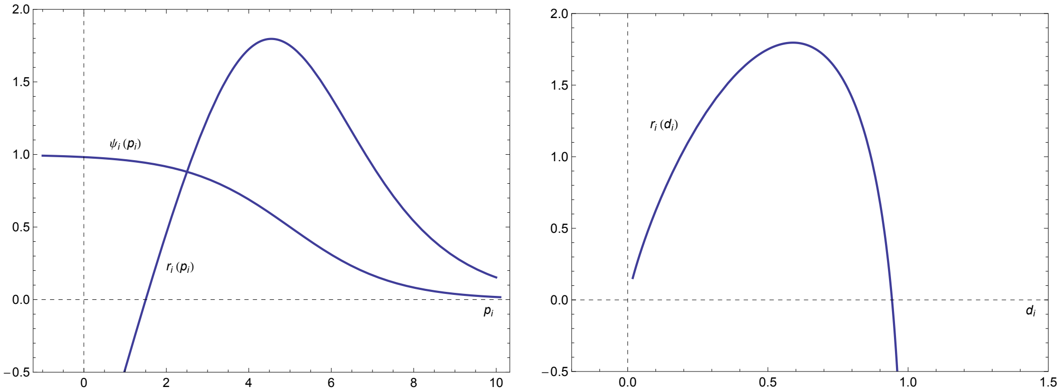


Figure 6: ***Left:*** Logistic demand $\psi_i(p_i)$ and revenue $r_i(p_i) = (p_i - c_i)\,\psi_i(p_i)$ as functions of price p_i. ***Right:*** Revenue $r(d_i) = \left(\psi_i^{-1}(d_i) - c_i\right) d_i$ as a function of demand d_i is concave for $d_i \in [0, 1]$.

a convex reformulation:

$$\begin{aligned} \underset{p_1,\ldots,p_n}{\text{maximize}} \quad & \sum_{i=1}^{n} (p_i - c_i)\,\psi_i(p_i) \\ \text{subject to} \quad & \sum_{i=1}^{n} \psi_i(p_i) \geq C. \end{aligned} \tag{81}$$

The model maximizes the total revenue objective for a cohort of n policies subject to a constraint on minimum retention level, where for i-th policyholder p_i is the proposed premium, $d_i = \psi_i(p_i)$ is the expected demand as a function of premium and c_i is the expected cost of claims. This optimization problem is not convex in general, however, for monotone demand functions from certain parametric families e.g. logistic and probit (see [50] for a detailed discussion) we can obtain a convex equivalent (or rather a maximization problem with a concave objective) parametrized by expected demand d_i (see Figure 6):

$$\begin{aligned} \underset{d_1,\ldots,d_n}{\text{maximize}} \quad & \sum_{i=1}^{n} \left(\psi_i^{-1}(d_i) - c_i\right) d_i \\ \text{subject to} \quad & 0 \leq d_i \leq 1, \ i = 1, \ldots, n \\ & \sum_{i=1}^{n} d_i \geq C. \end{aligned} \tag{82}$$

Here $\psi_i^{-1}(d_i)$ is the inverse demand function which can be uniquely defined for $d_i \in [0, 1]$.

It is worth noting that both claim costs and demand are usually not known precisely or even with reasonable accuracy and the rates obtained from (82) do not in fact maximize the expected revenue, but rather the "certainty equivalent" objective where we have replaced stochastic demand and claim costs with their expected values. A more accurate model would have the following form:

$$\begin{aligned} \underset{p_1,\ldots,p_n}{\text{maximize}} \quad & \mathbb{E}_\omega\left(\sum_{i=1}^{n} (p_i - c_i(\omega_i))\,\psi_i(p_i, \omega_i)\right) \\ \text{subject to} \quad & \mathbb{E}_\omega\left(\sum_{i=1}^{n} \psi_i(p_i, \omega_i)\right) \geq C, \end{aligned} \tag{83}$$

where the expectation is with respect to the joint distribution of individual demand functions and claims costs. Clearly some simplifying assumptions about the distribution will need to be made in order to make its estimation tractable. The interpretation of the volume constraint holding in expectation is also quite difficult. To address the latter point it may be appropriate to consider e.g.

the expected shortfall (reminiscent of conditional risk measures):

$$\mathbb{E}_\omega\Big(\big[C - \sum_{i=1}^{n} \psi_i(p_i, \omega_i)\big]_+\Big) \leq \epsilon. \tag{84}$$

Techniques that attempt to address this type of problems usually go under the name of stochastic optimization [47].

4 Solving convex optimization problems

Until recently, solving convex optimization problems required not inconsiderable subject matter expertise and even increasing availability of high quality open source and commercial solvers (sdpt3, sedumi, mosek, cvxopt) did not allay the situation. The reason for this is that these solvers require problems to be converted to one of several restrictive standard forms (e.g. a second order cone program). In some cases such a transformation is not possible and it may be necessary to develop a custom solver. For a potential user with a focus on applications these requirements are challenging to say the least.

This has recently changed with the introduction of cvx [20, 21], a high level specification language for general convex optimization. It provides a convenient interface for specifying convex optimization problems and then automates underlying mathematical steps for analysing and solving them. The underlying solvers can often handle sparse instances with millions of variables. While the idea of a an optimization modelling language is not new (e.g. gams), existing commercial offerings can not handle general convex problems.

Below is sample cvx/matlab code[8] for the problem (39):

```
n = length(y);  %length of timeseries y
m = n-2;        %length of Dw
lambda = 15;

I2 = speye(m,m);
O2 = zeros(m,1);
D = [I2 O2 O2]+[O2 -2*I2 O2]+[O2 O2 I2];

cvx_begin
   variable w(n)
   minimize(sum_squares(y-w)+lambda*norm(D*w,1))
cvx_end
```

Due to some limitations in the way the current version of cvx deals with problems that are not SDP representable (e.g. loss functions involving logarithms and exponentials) for large problems it may sometimes be necessary to implement an iterative reweighting scheme similar to Fisher scoring, such as in [37], with cvx in the inner loop. For very large problems decomposition methods are available, see [5] for an accessible introduction and examples.

[8] Please contact the author for larger examples described in this paper.

Appendix

A Basics of convex calculus

As mentioned earlier, it is often the easiest to verify convexity by checking whether a given function is composed from known convex primitives by applying convexity preserving transformations. Below we list a subset of these sufficient to verify convexity of problems described in this paper.

A.1 Functions of a single variable:

- *Exponential.* $e^{\alpha x}$ is convex on $\mathbb{R}$ for any $\alpha \in \mathbb{R}$.
- *Powers.* x^{α} is convex on $(0, +\infty)$ for $\alpha \geq 1$ or $\alpha \leq 0$, and concave for $0 \leq \alpha \leq 1$.
- *Logarithm.* $\log x$ is concave on $(0, +\infty)$.

A.2 Vector norms:

All vector norms on $\mathbb{R}^n$ are convex. A function $f : \mathbb{R}^n \to \mathbb{R}$ is called a norm if:

- $f(\mathbf{x}) \geq 0$ for all $\mathbf{x} \in \mathbb{R}^n$,
- $f(\mathbf{x}) = 0$ only if $\mathbf{x} = 0$,
- $f(\alpha\mathbf{x}) = |\alpha| f(\mathbf{x})$, for all $\mathbf{x} \in \mathbb{R}^n$ and $\alpha \in \mathbb{R}$,
- f satisfies the triangle inequality: $f(\mathbf{x} + \mathbf{y}) \leq f(\mathbf{x}) + f(\mathbf{y})$, for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$.

Such functions are usually denoted $\|\mathbf{x}\|$ rather than $f(\mathbf{x})$ with a subscript to indicate the exact norm being used. They can be interpreted as measuring the length of the elements of $\mathbb{R}^n$. Most common is the Euclidean or the ℓ_2-norm defined as:

$$\|\mathbf{x}\|_2 = \sqrt{\mathbf{x}^T\mathbf{x}} = \Big(\sum_i x_i^2\Big)^{\frac{1}{2}}. \tag{85}$$

Two other examples of norms on $\mathbb{R}^n$ are the absolute value or ℓ_1-norm:

$$\|\mathbf{x}\|_1 = \sum_i |x_i| \tag{86}$$

and the max or ℓ_∞-norm, given by:

$$\|\mathbf{x}\|_\infty = \max\left(|x_1|, \ldots, |x_n|\right). \tag{87}$$

These can be generalized as the ℓ_p-norm, for $p \geq 1$:

$$\|\mathbf{x}\|_p = \Big(\sum_i |x_i|^p\Big)^{\frac{1}{p}}, \tag{88}$$

where ℓ_∞-norm obtains as $p \to \infty$. It is easy to check convexity of norms from the definition.

A.3 Other functions on $\mathbb{R}^n$

- *Max function.* The function $\max(x_1,\ldots,x_n)$ is convex on $\mathbb{R}^n$. Note that it is distinct from the ℓ_∞-norm.
- *Soft max.* The so called "soft max" function $\log(e^{x_1}+\cdots+e^{x_n})$ is convex on $\mathbb{R}^n$. It can be viewed as a differentiable approximation of $\max(x_1,\ldots,x_n)$.
- *Geometric mean.* The geometric mean $f(\mathbf{x}) = (\prod_{i=1}^n x_i)^{\frac{1}{n}}$ is concave for $x_i > 0,\ i = 1,\ldots,n$.

A.4 Transformations that preserve convexity

Convexity is maintained for certain compositions of convex functions, for example:

A.4.1 Weighted sums:

A weighted sum of convex functions

$$f = \alpha_1 f_1 + \cdots + \alpha_n f_n \tag{89}$$

is convex if all the weights are non-negative, $\alpha_i \geq 0,\ i = 1,\ldots,n$. Similarly a weighted sum of concave functions is concave for non-negative weights.

A.4.2 Composition with an affine function:

If f is a convex function, $f:\mathbb{R}^m \to \mathbb{R}$, A is an $m \times n$ matrix and $\mathbf{b}$ a vector in $\mathbb{R}^n$ then $g:\mathbb{R}^n \to \mathbb{R}$:

$$g(\mathbf{x}) = f(A\mathbf{x} + \mathbf{b}) \tag{90}$$

is also a convex function. Similarly, g is concave if f is concave.

A.4.3 Composition of multivariable functions:

Suppose that a function $f:\mathbb{R}^n \to \mathbb{R}$ has the following form:

$$f(\mathbf{x}) = h(g(\mathbf{x})) = h(g_1(\mathbf{x}),\ldots,g_m(\mathbf{x})), \tag{91}$$

where $g_i:\mathbb{R}^n \to \mathbb{R},\ i = 1,\ldots,m$. Then the following holds:

- f is convex if functions g_i are convex and h is convex and non-decreasing in each argument,
- f is convex if g_i are concave and h is non-increasing in each argument.

B Some common GLMs as convex problems

B.1 Gaussian

Assume $y_i \in R$ follows a Gaussian distribution with known variance σ^2 and mean μ_i then its probability density function is:

$$p(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i-\mu_i)^2}{2\sigma^2}\right). \tag{92}$$

We can then express μ_i as a linear function (ignoring the intercept term for clarity) of the vector of explanatory variables $\mathbf{x}_i \in \mathbb{R}^m$, parametrized by $\mathbf{w} \in \mathbb{R}^m$:

$$\mu_i = \mathbf{w}^T \mathbf{x}_i. \tag{93}$$

The likelihood function for n independent observations $(y_i, \mathbf{x}_i)$ is given by:

$$p(\mathbf{y}; X\mathbf{w}) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i - \mathbf{w}^T \mathbf{x}_i)^2}{2\sigma^2}\right), \tag{94}$$

where X is the design matrix with rows $\mathbf{x}_i$ for $i = 1, \dots, n$ and the loss or negative log-likelihood function is as follows:

$$\begin{aligned}\mathcal{L}(\mathbf{y}; X\mathbf{w}) &= -\sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(y_i - \mathbf{w}^T \mathbf{x}_i)^2}{2\sigma^2}\right)\right) \\ &= \sum_{i=1}^{n} \frac{1}{2\sigma^2}(y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \frac{1}{2}\log(2\pi\sigma^2).\end{aligned} \tag{95}$$

We can obtain an estimate of w by solving a convex optimization problem (to check convexity in w, observe that it is a square of ℓ_2-norm composed with an affine function):

$$\underset{\mathbf{w}}{\text{minimize}} \quad \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2, \tag{96}$$

which in matrix notation can be restated as:

$$\underset{\mathbf{w}}{\text{minimize}} \quad \|\mathbf{y} - X\mathbf{w}\|_2^2. \tag{97}$$

B.2 Poisson

Take $y_i \in \mathbb{Z}^+$ to be a random variable with Poisson distribution and mean $\mu_i > 0$:

$$P(y_i = k) = \frac{e^{-\mu_i} \mu_i^k}{k!}. \tag{98}$$

The mean μ_i can be modelled via the log link as a linear function of the vector of explanatory variables $\mathbf{x}_i \in \mathbb{R}^m$:

$$\mu_i = \exp(\mathbf{w}^T \mathbf{x}_i). \tag{99}$$

Here $\mathbf{w} \in \mathbb{R}^m$ is the parameter vecor. Given n independent observations $(y_i, \mathbf{x_i}),\ i = 1, \dots, n$, the likelihood function is as follows:

$$p(\mathbf{y}; X\mathbf{w}) = \prod_{i=1}^{n} \frac{\exp\left(-\exp(\mathbf{w}^T \mathbf{x}_i)\right) \exp(\mathbf{w}^T \mathbf{x}_i)^{y_i}}{y_i!}, \tag{100}$$

with the corresponding loss or negative log-likelihood:

$$\mathcal{L}(\mathbf{y}; X\mathbf{w}) = \sum_{i=1}^{n} \left(\exp(\mathbf{w}^T \mathbf{x}_i) - y_i \mathbf{w}^T \mathbf{x}_i + \log(y_i!)\right). \tag{101}$$

A maximum likelihood estimate of w can then be obtained by solving the convex loss minimization problem (convexity follows directly from the composition rules defined in A.4):

$$\underset{\mathbf{w}}{\text{minimize}} \quad \sum_{i=1}^{n} \left(\exp(\mathbf{w}^T \mathbf{x}_i) - y_i \mathbf{w}^T \mathbf{x}_i \right). \tag{102}$$

B.3 Logistic

Consider a random variable $y_i \in \{0, 1\}$, with the binomial distribution:

$$P(y = 1) = \mu_i, \;\; P(y = 0) = 1 - \mu_i, \tag{103}$$

where $\mu_i \in [0, 1]$ is the expected value of y_i. It depends on a linear function of a vector of explanatory variables $\mathbf{x}_i \in \mathbb{R}^m$ via the logistic link:

$$\mu_i = \frac{\exp(\mathbf{w}^T \mathbf{x})}{1 + \exp(\mathbf{w}^T \mathbf{x})}, \tag{104}$$

where $\mathbf{w} \in \mathbb{R}^m$ is the vector of model parameters. Given n independent observations $(y_i, \mathbf{x}_i), i = 1, \ldots, n$ and defining $I_0 = \{i \mid y_i = 0\}$ and $I_1 = \{i \mid y_i = 1\}$ we obtain the following likelihood function:

$$p(\mathbf{y}; X\mathbf{w}) = \prod_{i \in I_1} \frac{\exp(\mathbf{w}^T \mathbf{x}_i)}{1 + \exp(\mathbf{w}^T \mathbf{x}_i)} \prod_{i \in I_0} \left(1 - \frac{\exp(\mathbf{w}^T \mathbf{x}_i)}{1 + \exp(\mathbf{w}^T \mathbf{x}_i)} \right). \tag{105}$$

The corresponding negative log-likelihood or loss function has the form:

$$\begin{aligned} \mathcal{L}(\mathbf{y}; X\mathbf{w}) &= - \sum_{i \in I_1} \log \left(\frac{\exp(\mathbf{w}^T \mathbf{x}_i)}{1 + \exp(\mathbf{w}^T \mathbf{x}_i)} \right) - \sum_{i \in I_0} \log \left(1 - \frac{\exp(\mathbf{w}^T \mathbf{x}_i)}{1 + \exp(\mathbf{w}^T \mathbf{x}_i)} \right) \\ &= - \sum_{i \in I_1} \mathbf{w}^T \mathbf{x}_i + \sum_{i=1}^{n} \log \left(1 + \exp(w^T \mathbf{x}_i) \right) \end{aligned} \tag{106}$$

It is easy to check that $\mathcal{L}(\mathbf{y}; X\mathbf{w})$ is a convex function in w (composition of soft max and affine). Maximum likelihood estimation for logistic regression is then equivalent to a convex loss minimization problem:

$$\underset{\mathbf{w}}{\text{minimize}} \quad - \sum_{i \in I_1} \mathbf{w}^T \mathbf{x}_i + \sum_{i=1}^{n} \log \left(1 + \exp(\mathbf{w}^T \mathbf{x}_i) \right). \tag{107}$$

References

[1] K. Antonio and J. Beirlant. Actuarial statistics with generalized linear mixed models. *Insurance: Mathematics and Economics*, 40:58–76, 2007.

[2] D. Bertsekas. *Nonlinear programming*. Athena Scientific Belmont, 1999.

[3] D Bertsimas and J Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific, 1st edition, 1997.

[4] G. Bohlmann. Ein Ausgleichungsproblem. *Nachrichten Gesellschaft Wissenschaften Gottingen. Math.-Phys. Klasse*, pages 260–271, 1899.

[5] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3:1–122, 2011.

[6] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Presss, 2004.

[7] P. Brockett. Information theoretic approach to actuarial science: a unification and extension of relevant theory and applications. *Transactions of the Society of Actuaries*, 43:73–114, 1991.

[8] L. Brown. *Fundamentals of statistical exponential families with applications in statistical decision theory*. Institute of Mathematical Statistics, Hayward, CA, 1986.

[9] E. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Transactions on Information Theory*, 2006.

[10] F. Y. Chan, L. K. Chan, J. Falkenberg, and M. H. Yu. Applications of linear and quadratic programming to some cases of the Whittaker-Henderson graduation method. *Scandinavian Actuarial Journal*, pages 141–153, 1986.

[11] F. Y. Chan, L. K. Chan, and M. H. Yu. A generalization of Whittaker-Henderson graduation. *Transactions of Society of Actuaries*, 36:183–211, 1984.

[12] S. Chen, D. Donoho, and M. Saunders. Atomic decomposition by basis pursuit. *SIAM Review*, 2001.

[13] D. K. Dey, A. E. Gelfand, and F. Peng. Overdispersed generalized linear models. *Journal of Statistical Planning and Inference*, 64:93–108, 1997.

[14] D. Donoho. Compressed sensing. *IEEE Transactions on Information Theory*, 2006.

[15] M. Ezekiel. A method of handling curvilinear correlation for any number of variables. *Journal of the American Statistical Association*, 19:431–453, 1924.

[16] L. Fahrmeir. Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalised linear models. *Journal of the American Statistical Association*, 87:501–509, 1992.

[17] L. Fahrmeir and H. Kaufmann. On Kalman filtering, posterior mode estimation and Fisher scoring in dynamic exponential family regression. *Metrika*, 38:37–60, 1991.

[18] E. Frees, P. Shi, and E. Valdez. Actuarial applications of a hierarchical insurance claims model. *ASTIN Bulletin*, 39:165–197, 2009.

[19] A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press, 2006.

[20] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control*. 2008.

[21] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21, September 2010.

[22] T. Hastie and R. Tibshirani. Generalized additive models. *Statistical Science*, 1:297–318, 1986.

[23] T. Hastie and R. Tibshirani. *Generalized aditive models*. Chapman & Hall/CRC, 1990.

[24] R. Hodrick and E. Prescott. Postwar U.S. business cycles: An empirical investigation. *Journal of Money, Credit, and Banking*, 29:1–16, 1997.

[25] M. D. Hoffman and A. Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, In press.

[26] P. Huber. *Robust Statistics*. John Wiley & Sons, 1981.

[27] W. Jewell. The use of collateral data in credibility theory: a hierarchical model. *Giornale dell'Instituto Italiano degli Attuari*, 38:1–16, 1975.

[28] D. Jones and H. Gerber. Credibility formulas of the updating type. *Transactions of the Society of Actuaries*, 27:31–46, 1975.

[29] P. De Jong and B. Zehnwirth. Claims reserving state space models and the Kalman filter. *Journal of the Institute of Actuaries*, 110:157–181, 1983.

[30] P. De Jong and B. Zehnwirth. Credibility theory and the Kalman filter. *Insurance: Mathematics and Economics*, 2:281–286, 1983.

[31] B. Jørgensen. Exponential dispersion models (with discussion). *Journal of the Royal Statistics Society Series B*, 49:127–162, 1987.

[32] R. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960.

[33] S.J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. ℓ_1-trend filtering. *SIAM Review*, 51:339–360, 2009.

[34] R. Koenker. *Quantile regression*. Cambridge University Press, 2005.

[35] R. Koenker, P. Ng, and S. Portnoy. Quantile smoothing splines. *Biometrika*, 81:673–80, 1994.

[36] A. Kudryavtsev. Using quantile regression for rate–making. *Insurance: Mathematics and Economics*, 45:296–304, 2009.

[37] S. Lee, H. Lee, P. Abeel, and A. Ng. Efficient ℓ_1-regularized logistic regression. In *AAAI*, 2006.

[38] W. B. Lowrie. Multidimensional Whittaker-Henderson graduation with constraints and mixed differences. *Transactions of Society of Actuaries*, 45:215–255, 1993.

[39] R. Mehra. Credibility theory and kalman filtering with extensions. Technical Report RM 75-64, International Institute for Applied Systems Analysis, Schloss Laxenburg, Austria, 1975.

[40] H. Miller and P. Mulquiney. Credibility, penalised regression and boosting; let's call the whole thing off. In *Pricing and Underwriting Seminar*. Casualty Actuarial Society, 2011.

[41] J. Nelder and R. Verrall. Credibility theory and generalized linear models. *ASTIN Bulletin*, pages 71–82, 1997.

[42] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course (Applied Optimization)*. Springer, 1st edition, 2003.

[43] C.E. Rasmussen and C.K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2005.

[44] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[45] D. Schuette. A linear programming approach to graduation. *Transactions of Society of Actuaries*, 30:407–445, 1978.

[46] H. L. Seal. Graduation by piecewise cubic polynomials: A historical review. *Blätter der DGVFM*, 15:89–114, 1981.

[47] A. Shapiro, D. Dentcheva, and A. Ruszczyski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2009.

[48] M. L. Stein. *Interpolation of spatial data: Some theory for Kriging*. Springer-Verlag, New York, 1999.

[49] C. A. Sutton and A. McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4:267–373, 2012.

[50] K. Talluri and G. van Ryzin. *The Theory and Practice of Revenue Management*. Springer, 2004.

[51] G. Taylor. A Bayesian interpretation of Whittaker-Henderson graduation. *Insurance: Mathematics and Economics*, 11:7–16, 1992.

[52] G. Taylor. Technical aspects of domestic lines pricing. Technical report, University of Melbourne, Centre for Actuarial Studies, 1997.

[53] G. Taylor. Geographic premium rating by Whittaker spacial smoothing. *ASTIN Bulletin*, 31:147–160, 2001.

[54] G. Taylor. Second-order Bayesian revision of a generalised linear model. *Scandinavian Actuarial Journal*, pages 202–242, 2008.

[55] H. Taylor, S. Banks, and J. McCoy. Deconvolution with the ℓ_1 norm. *Geophysics*, 1979.

[56] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 1996.

[57] R. Tibshirani, M. Saunders, S. Rosset, J Zhu, and K. Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistics Society Series B*, 67:91–108, 2005.

[58] A. Tychonoff and V. Arsenin. *Solution of Ill-posed Problems*. Winston & Sons, 1977.

[59] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, 1995.

[60] R. Verall. A state space formulation of Whittaker graduation with existensions. *Insurance: Mathematics and Economics*, 13:1–14, 1993.

[61] R. V. Vohra. *Advanced Mathematical Economics*. Routledge, London and New York, 2005.

[62] M. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

[63] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *ICML*, 2011.

[64] E. Whittaker. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:63–75, 1923.

[65] M. Wright. The interior-point revolution in optimization: history, recent developments, and lasting consequences. *Bulletin of the American Mathematical Society*, 42:39–56, 2005.

[66] J. Yan, J. Guszcza, M. Flynn, and C.-S. P. Wu. Applications of the offset in property-casualty predictive modeling. In *Forum*. Casualty Actuarial Society, 2009.

[67] M. Yuan and L. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society*, 2006.

[68] H. Zou and T. Hastie. Regularization and variable selection via the elastic net,. *Journal of the Royal Statistical Society*, 2005.