general insurance seminar

Tides of Change

12-13 November 2012
Sofitel Sydney Wentworth

Actuaries Institute

A convex optimisation perspective on dynamic GLMs with applications to automated portfolio monitoring

Prepared by D. Semenovich and M. McLean

Presented to the Actuaries Institute
General Insurance Seminar
12 – 13 November 2012
Sydney

Institute of Actuaries of Australia
ABN 69 000 423 656
Level 7, 4 Martin Place, Sydney NSW Australia 2000
t +61 (0) 2 9233 3466 **f** +61 (0) 2 9233 3446
e actuaries@actuaries.asn.au **w** www.actuaries.asn.au

A convex optimisation perspective on dynamic GLMs with applications to automated portfolio monitoring

Dimitri Semenovich[1], **Michael McLean**[2]
[1]UNSW, [2]Finity

Abstract: Many insurance problems require fitting of statistical models. Updating these models to reflect new experience usually takes the form of periodic manual reviews, often without a formal process to evaluate the credibility of resulting parameter changes.

In the present paper we describe a framework that allows automatic updating of parameters of a wide range of models while respecting certain optimality properties. Special cases include dynamic GLMs, the Kalman filter and a number of classical credibility and time series models. Formulating parameter estimation as convex optimisation problems makes it is easy to introduce additional constraints on parameter ranges, ensuring robustness of the procedure in a production environment. Applications include "live" monitoring of conversion rates, claims experience, changes in exposure and demand. It is also hoped that this presentation can clarify some of the intuition behind classical methods.

Keywords: Dynamic generalised linear models, portfolio monitoring, credibility, graduation, convex optimisation.

Acknowledgement: Some of the proposed applications are due to Charles Pollack, whom the authors thank for the helpful discussion.

1 Introduction

The past twenty years have seen dramatic changes in general insurance premium rating methodologies. First generalised linear models (GLMs) and later techniques from computational statistics and machine learning[1] such as "random forests" and "gradient boosting" [HTF08] have been adopted for personal lines pricing in most deregulated markets. Classical actuarial tools such as graduation and credibility theory [BG05] have not kept pace with the new practices and yet they deal with important premium rating issues, such as optimal parameter updates, which cannot be readily addressed in the standard GLM framework.

In this paper we argue that by focusing on the underlying optimisation problems it is possible not only to unify credibility and related methods with GLMs but to significantly extend the practical reach of both.

1.1 Overview of mathematical optimisation

Broadly, a substantial number of inference problems in statistics can be reduced to solving specific optimisation problems.

Within optimisation, it is especially productive to focus on models for which there exist numerical techniques that are guaranteed to find a global solution in polynomial time. The two large

[1]A variant of computational statistics practiced in computer science departments, with primary focus on empirical out of sample performance and exploratory data analysis.

classes of such computationally tractable problems are convex problems[2] and those that can be reduced to finding eigenvalues or singular values of certain matrices, e.g. the so called S-procedure.

Recent advances in algorithms [Wri05] and modelling systems for convex optimisation have dramatically increased their applicability. Today it is possible to use the same software to effectively solve robust versions of the classical Markowitz portfolio problem, maximum likelihood estimation in generalised linear models and perform image denoising.

The key practical benefit of working directly with modelling systems such as CVX [GB10] and numerical solvers rather than standard statistical software lies in the ability to modify the models to reflect specific aspects of judgment and business requirements (e.g., some parameters should be always positive). In the framework of credibility theory, for example, even a slight extension effectively turns into a research problem (witness the proliferation of named models in the 1970s), where one may or may not finally succeed. On the other hand familiarity with the basics of convex analysis [BV04, Roc70] and a few heuristics will permit effective creation of custom models. 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. In particular, all the optimisation problems mentioned in this paper are convex, which can be easily verified using "convex calculus" as in [BV04, Chapter 3].

1.2 Insurance applications of dynamic models

Models are at the heart of many insurance problems. Updating these models to reflect new experience usually takes the form of periodic manual reviews but often without a formal process to evaluate the credibility of resulting parameter changes. The situation is paralleled by various regularly generated internal reports which often are just one or two way tables containing comparisons against budget or year-on-year.

Finding actionable trends in data when viewed in this format can be challenging at best. While the issue has received surprisingly little attention until recently, the need for more sophisticated monitoring has been pointed out by Taylor [Tay11] and Berry et al. [BHMM09]. Indeed approaches to model evaluation described in the latter paper are complimentary to the parameter update strategies described here.

In actuarial practice dynamic model revision has traditionally been addressed in the setting of credibility theory [BG05]. The associated literature is somewhat difficult, suffering from overly abstract notation and at the same time understating the domain of its own applicability due to a strong focus on (linear) closed form solutions.

In this paper we describe a framework that allows automatic updating of parameters for a wide range of models. Special cases include dynamic GLMs, Kalman filter and a number of classical credibility formulas. As already mentioned, posing parameter estimation as a convex optimisation problem makes it is easy to introduce additional constraints on parameter ranges, ensuring robustness of the procedure in a production environment.

Applications include "live" monitoring of conversion rates, claims process and costs, changes in exposure and demand. It is also hoped that the new presentation can clarify some of the intuition behind classical methods.

[2] 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. It is worth noting that one of many fields where convex analysis has proven to be the key tool is mathematical economics; see [Voh05] for an accessible introduction from this point of view.

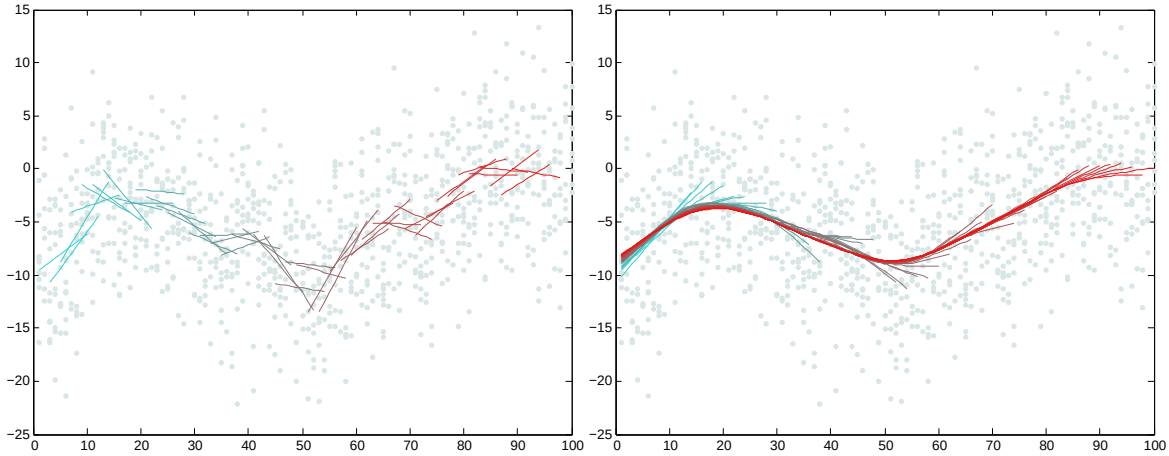

Figure 1: Best viewed in colour. *Left:* Collection of linear models fit locally as new data (denoted by grey dots) becomes available. *Right:* Estimates of the current trend obtained from a dynamic model with second order smoothing. Note that at each time step the entire history is re-estimated.

2 Dynamic parameter estimation

To illustrate the key points and introduce notation we discuss the simple case of mean estimation when the underlying distribution evolves over time. Initially we consider only the so called *linear*[3] estimators, which coincide with maximum likelihood estimators for the Gaussian distribution but otherwise may be less efficient. We assume that there are m time periods and in the time period t we obtain n new observations, with i-th observation at time t denoted as y_{ti}. The simplest approach is to compute the sample average, updating it to reflect new data as they come in. It is well known that the sample average:

$$w^* = \frac{1}{mn} \sum_{t=1}^{m} \sum_{i=1}^{n} y_{ti} \tag{1}$$

solves the following quadratic loss minimisation problem:

$$\underset{w}{\text{minimise}} \sum_{t=1}^{m} \|\mathbf{y}_t - \mathbf{1}w\|_2^2 = \sum_{t=1}^{m} \sum_{i=1}^{n} (y_{ti} - w)^2. \tag{2}$$

Alternatively we can calculate the sample means of all m time periods independently:

$$\underset{\mathbf{w}}{\text{minimise}} \sum_{t=1}^{m} \|\mathbf{y}_t - \mathbf{1}w_t\|_2^2 = \sum_{t=1}^{m} \sum_{i=1}^{n} (y_{ti} - w_t)^2, \tag{3}$$

with the following analytic solutions:

$$w_t^* = \frac{1}{n} \sum_{i=1}^{n} y_{ti}. \tag{4}$$

Both methods are clearly somewhat unsatisfying. In the first case we effectively assume that the mean stays constant over time and the sensitivity of the estimate to new observations diminishes as new data are acquired, while in the second case the changes in the mean are assumed to be uncorrelated and we forgo any sharing of information between adjoining time periods.

One compromise might be a form of local regression where we fit a linear model over a time

[3] Linear estimators, or at least *optimal* linear estimators, actually minimise a quadratic objective with the name referring to the fact that such problems have analytic solutions in the form of linear functions of the original problem data.

window of fixed width $q+1$ to form the estimate:

$$\underset{w,b}{\text{minimise}} \sum_{t=m-q}^{m} \|\mathbf{y}_t - \mathbf{1}(tw+b)\|_2^2 = \sum_{t=m-q}^{m} \sum_{i=1}^{n} (y_{ti} - (tw+b))^2, \tag{5}$$

which, unlike (2), provides an indication of the current trend.

Another way to approach the problem is to observe that (2) can be equivalently rewritten in the form similar to (3) by introducing some equality constraints:

$$\begin{aligned} \underset{\mathbf{w}}{\text{minimise}} \quad & \sum_{t=1}^{m} \|\mathbf{y}_t - \mathbf{1}w_t\|_2^2 = \sum_{t=1}^{m} \sum_{i=1}^{n} (y_{ti} - w_t)^2 \\ \text{subject to} \quad & w_{t+1} - w_t = 0, \;\; t = 1, \dots, m-1. \end{aligned} \tag{6}$$

It is then natural to replace hard constraints with e.g. a quadratic penalty term (absolute values or even maximum of differences etc. can also be considered[4]):

$$\underset{\mathbf{w}}{\text{minimise}} \sum_{t=1}^{m} \|\mathbf{y}_t - \mathbf{1}w_t\|_2^2 + \lambda \|D^{(1,m)}\mathbf{w}\|_2^2 = \sum_{t=1}^{m} \sum_{i=1}^{n} (y_{ti} - w_t)^2 + \lambda \sum_{t=1}^{m-1} (w_{t+1} - w_t)^2, \tag{7}$$

delivering an effective compromise solution, where we can "dial" between (2) and (3) by choosing the value of $\lambda \geq 0$. Here $D^{(1,m)}$ is the $(m-1) \times m$ first order finite differences matrix:

$$D^{(1,n)} = \begin{bmatrix} -1 & 1 & & & \\ & -1 & 1 & & \\ & & \dots & & \\ & & & -1 & 1 \end{bmatrix} \tag{8}$$

and the k-th order finite differences matrix $D^{(k,m)} \in \mathbb{R}^{(m-k)\times m}$ is recursively defined as:

$$D^{(k,m)} = D^{(1,m-k)} D^{(k-1,m)}, \; k = 2, 3, \dots \tag{9}$$

Heuristically, the "fidelity" term $\sum_{t=1}^{m} \|\mathbf{y}_t - \mathbf{1}w_t\|_2^2$ encourages the solution $\mathbf{w}$ to be close to the original data $\mathbf{y}_t$ and the smoothness or regularisation term $\|D^{(1,m)}\mathbf{w}\|_2^2$ penalises non-zero entries of $D^{(1,m)}\mathbf{w}$ (the discretised first derivative) of $\mathbf{w}$. If we would like to allow linear or higher order polynomial trends without penalty, we can instead use $\|D^{(k,m)}\mathbf{w}\|_2^2$ with $k = 2$ or $k \geq 3$, corresponding to discretised second and higher derivatives respectively.

It is important to consider what happens as we obtain data for the new time period $m+1$. Perhaps the cleanest way to approach this is to form a new optimisation problem (e.g. 7) with updated data and obtain a new vector of estimates for the entire history of the process $\mathbf{w}^* = [w_1^*, \dots, w_m^*, w_{m+1}^*]^T$. Historically a number of schemes to efficiently compute w_{m+1}^* given the estimates at time m have been proposed but advances in computing power have rendered them considerably less relevant.

Figure 1 shows the results of running local regression (5) and mean estimation penalised with second order differences (7) on some syntethic data.

2.1 Exponential families and quantile splines

The same idea as in the previous section can also be applied to maximum likelihood estimation. For example, consider a single parameter exponential family of distributions (see Appendix B for

[4] There exists a considerable arsenal of justifications for these heuristics, e.g. Gaussian or Laplacian priors, optimal estimators under quadratic loss, sparsity properties of the solution, robustness considerations etc.

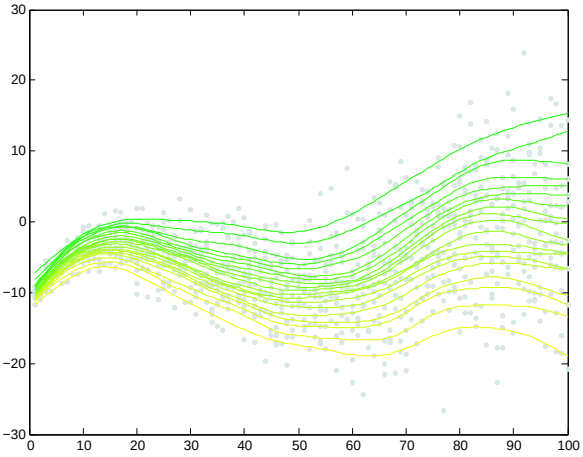

Figure 2: Quantile splines (12) applied to estimate 20 equally spaced distribution percentiles (denoted by colour gradient) given the information at $t = 100$. Best viewed in colour.

details on notation) where the parameter can change over time:

$$p(y|w_t) = h_0(y) \exp\big(w_t \phi(y) - A(w_t)\big). \tag{10}$$

We can attain similar results to (7) by solving the following penalised maximum likelihood problem:

$$\underset{\mathbf{w}}{\text{minimise}} \quad \sum_{t=1}^{m} \sum_{i=1}^{n} \big(A(w_t) - w_t \phi(y_{ti})\big) + \lambda \|D^{(k,m)} \mathbf{w}\|_2^2. \tag{11}$$

This formulation can also be extended to exponential families with multiple parameters (e.g. heteroscedastic multivariate Gaussian), see [WBAW12] for a recent example.

An interesting alternative is to estimate distribution quantiles directly, instead of assuming a parametric form and then trying to find the parameters. This can be accomplished by replacing quadratic loss in (7) with the quantile loss (see Appendix A), so that solving the following problem estimates the τ-th quantile of the distribution given data up to time m:

$$\underset{\mathbf{w}}{\text{minimise}} \quad \sum_{t=1}^{m} \sum_{i=1}^{n} \rho_\tau(y_{ti} - w_t) + \lambda \|D^{(k,m)} \mathbf{w}\|_2^2, \tag{12}$$

yielding a variant of the so called *quantile splines* [KNP94]. Figure 2 shows an example of quantile splines applied to estimate a time varying distribution (at time $t = 100$).

2.2 Relation to classical actuarial models

2.2.1 Graduation

In actuarial literature a variant of (7) was first proposed by Bohlmann [Boh99, Sea81] at the end of the 19th century in the context of mortality graduation. The goal of graduation is to smooth a suitably normalised sequence of death counts $\mathbf{y} = [y_1, y_2, \dots, y_m]^T$ indexed by age at death. The differences from (7) are that the indexing variable is age rather than time, there is a single observation per index value and it is generally not expected that the indexing set grows (i.e. no new data are added):

$$\underset{\mathbf{w}}{\text{minimise}} \ \|\mathbf{y} - \mathbf{w}\|_2^2 + \lambda \|D^{(k,m)} \mathbf{w}\|_2^2, \tag{13}$$

with $\lambda \geq 0$ and $k = 1$. Whittaker [Whi23] described the underlying probabilistic model and an approximate solution method for the case of third order differences and weighted fidelity term. In his remarkable paper Schuette [Sch78] proposed the formulation using ℓ_1-norm penal-

ties (weights omitted to simplify presentation):

$$\underset{\mathbf{w}}{\text{minimise}}\ \|\mathbf{y}-\mathbf{w}\|_1+\lambda\|D^{(k,m)}\mathbf{w}\|_1, \tag{14}$$

After applying a standard transformation this optimisation problem can be reformulated as a linear program. Chan et al. [CCFY86] show that for $p,q\in\{1,2,\infty\}$ the mixed ℓ_p and ℓ_q norm graduation problem:

$$\underset{\mathbf{w}}{\text{minimise}}\ \|\mathbf{y}-\mathbf{w}\|_p+\lambda\|D^{(k,m)}\mathbf{w}\|_q, \tag{15}$$

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 [HP97]. More recently, a variant of (15):

$$\underset{\mathbf{w}}{\text{minimise}}\ \|\mathbf{y}-\mathbf{w}\|_2^2+\lambda\|D\mathbf{w}\|_1, \tag{16}$$

with $p=2$ (or equivalently squared ℓ_2-norm), $q=1$ and $D=D^{(1,m)}$ has been popularised in applied statistics literature as *fused lasso* [TSR+05]. In signal processing the same formulation is called *total variation denoising*. This procedure usually gives a piecewise constant solutions $\mathbf{w}^*$ i.e. discretised first derivative $D^{(1,m)}\mathbf{w}^*$ has mostly zero entries due to "sparsity inducing" property of ℓ_1-norm penalties (also pointed out in the discussion of [Sch78]). Similarly, using second order differences $D^{(2,m)}$ often results in a piecewise linear $\mathbf{w}^*$ and has been described as ℓ_1 *trend filtering* [KKBG09] and *quantile splines* [KNP94], the latter (12) replacing the quadratic term with quantile loss. All of these considerations can, of course, be directly applied to modify the problem (7).

2.2.2 Credibility

Another classical actuarial model essentially identical to the above setting is the Jones and Gerber credibility formula [JG75, BG05]. Below are its simplified assumptions (omitting the conditions on w_1):

$$\begin{aligned}&\mathbb{E}(w_{t+1}-w_t)=0, &&\operatorname{Var}(w_{t+1}-w_t)=\sigma_w^2,\\ &\mathbb{E}(y_t|w_t)=w_t, &&\operatorname{Var}(y_t|w_t)=\sigma_y^2.\end{aligned} \tag{17}$$

Best linear unbiased predictor of the parameter vector $\mathbf{w}$ is then obtained as a solution of:

$$\underset{\mathbf{w}}{\text{minimise}}\ \frac{1}{\sigma_y^2}\sum_{t=1}^{m}\sum_{i=1}^{n}(y_{ti}-w_t)^2+\frac{1}{\sigma_w^2}\sum_{t=1}^{m-1}(w_t-w_{t+1})^2 \tag{18}$$

or more concisely in vector notation:

$$\underset{\mathbf{w}}{\text{minimise}}\ \frac{1}{\sigma_y^2}\sum_{t=1}^{m}\|\mathbf{y}_t-\mathbf{1}w_t\|_2^2+\frac{1}{\sigma_w^2}\|D^{(1,m)}\mathbf{w}\|_2^2. \tag{19}$$

When there is only a single observation per time step, this is effectively identical to Whittaker graduation with first order differences as in (13).

It seems that a particularly effective way to come to terms with the classical credibility literature is by first examining the connections with the Kalman filter [Meh75, JZ83b] and then explicitly writing out the underlying optimisation problems following the approach in the next section.

3 Kalman filter and dynamic GLMs

Kalman filter [Kal60] and related ideas have played a central role in the success 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 [MS85], quickly followed by adoption for a wide range of aerospace problems [GA10]. In these applications the goal is typically to track the "state" of a missile or a spacecraft following Newtonian dynamics. The state vector would contain the current position, 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 [Meh75, JZ83a, JZ83b, Tay08], particularly in claims reserving. 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.

3.1 Kalman filter as an optimisation problem

Kalman filter can be productively conceptualised as a "dynamic" extension of the standard least squares problem:

$$\underset{\mathbf{w}}{\text{minimise}} \ \|\mathbf{y} - X\mathbf{w}\|_2^2, \tag{20}$$

Following (6) we partition the design matrix X and the response vector $\mathbf{y}$ into m row blocks (corresponding to time periods):

$$X = \begin{bmatrix} X_1 \\ \cdots \\ X_m \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \cdots \\ \mathbf{y}_m \end{bmatrix}. \tag{21}$$

The least squares problem (20) can then be transformed into an equivalent problem with m copies of the parameter vector by introducing some equality constraints:

$$\begin{aligned} \underset{\mathbf{w}_1,\dots,\mathbf{w}_m}{\text{minimise}} \quad & \sum_{t=1}^{m} \|\mathbf{y}_t - X_t\mathbf{w}_t\|_2^2 \\ \text{subject to} \quad & \mathbf{w}_{t+1} - \mathbf{w}_{t+1} = \mathbf{0}, \ \ t = 1, \dots, m-1. \end{aligned} \tag{22}$$

The underlying probability model can be written in the state space form with the identity state transition matrix, no state transition noise and i.i.d. Gaussian observation noise[5], where $\mathbf{w}_t$ is the unobserved state vector and $\mathbf{y}_t$ are the observations associated with time dependent observation matrices X_t:

$$\begin{aligned} &\mathbf{w}_{t+1} = \mathbf{w}_t, \qquad && \mathbf{y}_t = X_t\mathbf{w}_t + \epsilon_t, \\ & && \epsilon_t \sim \mathcal{N}(0, I). \end{aligned} \tag{23}$$

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

$$\begin{aligned} &\mathbf{w}_{t+1} = \mathbf{w}_t + \nu_t, \qquad && \mathbf{y}_t = X_t\mathbf{w}_t + \epsilon_t, \\ &\nu_t \sim \mathcal{N}(0, I), && \epsilon_t \sim \mathcal{N}(0, I) \end{aligned} \tag{24}$$

[5] We can in fact avoid the Gaussianity assumption by posing a quadratic loss function instead which yields equivalent estimators (the approach taken in section 2.1).

we effectively relax the equality constraints in (22) replacing them with a squared ℓ_2-norm penalty. It is then possible to perform estimation by solving the following convex optimisation problem:

$$\underset{\mathbf{w}_1,\dots,\mathbf{w}_m}{\text{minimise}} \ \sum_{t=1}^{m} \|\mathbf{y}_t - X_t\mathbf{w}_t\|_2^2 + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2, \tag{25}$$

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

$$\underset{\mathbf{w}}{\text{minimise}} \ \|\mathbf{y}' - X'\mathbf{w}\|_2^2, \tag{26}$$

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

$$X' = \begin{bmatrix} X_1 & & & & \\ -I & I & & & \\ & & \cdots & & \\ & & & -I & I \\ & & & & X_m \end{bmatrix}, \qquad \mathbf{y}' = \begin{bmatrix} \mathbf{y}_1 \\ 0 \\ \cdots \\ 0 \\ \mathbf{y}_m \end{bmatrix}, \qquad \mathbf{w} = \begin{bmatrix} \mathbf{w}_1 \\ \cdots \\ \mathbf{w}_m \end{bmatrix}. \tag{27}$$

The objective in (25) simultaneously performs both "filtering" and "smoothing" conditional on all the observations up to time m. If new information becomes available the augmented optimisation problem should be solved again to obtain new estimates of the entire history of state transitions $\mathbf{w}^* = [\mathbf{w}_1^*, \dots, \mathbf{w}_m^*, \mathbf{w}_{m+1}^*]^T$. 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 "smoothing" pass to update estimates of past states $\mathbf{w}_1, \dots, \mathbf{w}_m$.

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 (26) 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:

$$\begin{aligned} \mathbf{w}_{t+1} &= F\mathbf{w}_t + \nu_t, & \mathbf{y}_t &= X_t\mathbf{w}_t + \epsilon_t, \\ \nu_t &\sim \mathcal{N}(0, \Sigma_\nu), & \epsilon_t &\sim \mathcal{N}(0, \Sigma_\epsilon) \end{aligned} \tag{28}$$

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:

$$\underset{\mathbf{w}_1,\dots,\mathbf{w}_m}{\text{minimise}} \ \sum_{t=1}^{m} \|\mathbf{y}_t - X_t\mathbf{w}_t\|_{\Sigma_\epsilon^{-1}}^2 + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - F\mathbf{w}_t\|_{\Sigma_\nu^{-1}}^2. \tag{29}$$

We can easily recover both Whittaker graduation and Jones-Gerber credibility as state space models by choosing one a dimensional state vector w_t with the observation matrix X_t a constant vector $\mathbf{1}$ [JZ83b, Ver93].

3.2 Some extensions to the dynamic models

We can use any convex loss for the observations, such as quantile loss (Appendix A) or the negative log-likelihood for conditional exponential families (Appendix C), and still end up with a convex

optimisation problem:

$$\underset{\mathbf{w}_1,\dots,\mathbf{w}_m}{\text{minimise}} \quad \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t\mathbf{w}_t) + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2. \tag{30}$$

In the context of dynamic generalised linear models this corresponds to posterior mode estimation as proposed by Fahrmeier [FK91, Fah92]. Non-Gaussian state noise is another possibility. It may be beneficial to apply ℓ_1-norm penalty to state changes provided most of the time parameters stay constant with occasional large jumps:

$$\underset{\mathbf{w}_1,\dots,\mathbf{w}_m}{\text{minimise}} \quad \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t\mathbf{w}_t) + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_1. \tag{31}$$

Another possibility is a combination of norms. For example 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:

$$\underset{\mathbf{w},\mathbf{c}}{\text{minimise}} \quad \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t(\mathbf{w}_t + \mathbf{c}_t)) + \lambda \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_1 + \mu \sum_{t=1}^{m-1} \|\mathbf{c}_{t+1} - \mathbf{c}_t\|_2^2. \tag{32}$$

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

$$\underset{\mathbf{w}_1,\dots,\mathbf{w}_m}{\text{minimise}} \quad \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t\mathbf{w}_t) + \sum_{t=1}^{m-2} \|\mathbf{w}_{t+2} - 2\mathbf{w}_{t+1} + \mathbf{w}_t\|_1. \tag{33}$$

It is also possible to add general convex inequality and linear equality constraints on the parameters. For example seasonality adjustments can be handled by introducing new variables $\mathbf{c}_1, \dots, \mathbf{c}_m$ and equality constraints:

$$\begin{aligned} \underset{\mathbf{w},\mathbf{c}}{\text{minimise}} \quad & \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t(\mathbf{w}_t + \mathbf{c}_t)) + \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2 \\ \text{subject to} \quad & \mathbf{c}_t = \mathbf{c}_{t+k}, \;\; t = 1, \dots, m-k \\ & \sum_{t=1}^{m} \mathbf{c}_t = 0. \end{aligned} \tag{34}$$

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

4 Insurance applications

It has been often argued that actuaries have the data and the tools to materially improve the performance of their companies by becoming involved in a wide range of operational decisions beyond the traditional pricing and reserving functions. Some of the "actions" available to an insurer include its claims management strategy, pricing and profit loadings, denial of cover, policy features and exclusions, and to some extent targeted advertising. Below we describe how the methods developed in this paper can be applied to inform decision making in several of these areas.

4.1 Conversion rates

Estimation of conversion rates is one application where the model parameters can be expected to rapidly evolve in response to changes in the market position, advertising initiatives etc. Improved market transparency can only exacerbate these effects. While we do not discuss price testing directly, it is often best considered in the same framework.

The problem is modelled well by a dynamic GLM with logistic loss and a penalty on the first order differences in parameters, allowing for shifts in the profile of conversion probabilities:

$$\underset{\mathbf{w}}{\text{minimise}} \quad \sum_{t=1}^{m} \sum_{i=1}^{n} \Big(\log\left(1 + \exp(\mathbf{w}_t^T \mathbf{x}_{ti})\right) - y_{ti} \mathbf{w}_t^T \mathbf{x}_{ti} \Big) + \lambda \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_2^2. \tag{35}$$

Appropriate time step may be on the order of one day to one week, depending on the distribution channel, with the smoothing parameter λ chosen by cross validation.

Sudden changes in conversion rates could potentially trigger an investigation into the profitability of the affected segments, in particular if the affected segments are not well represented historically.

4.2 Changes in exposure

Monitoring conversions alone is in a sense insufficient as a "conversion" is conditional on a renewal notice or a quote having been issued in the first place. Looking at changes in exposure directly can thus be an important additional tool to confirm the impact of rate changes, marketing initiatives and to keep track of possible deterioration of the known cross-subsidies.

We can approach this by fitting multivariate distributions $p(\mathbf{x}; \theta_t)$ to the risk factors $\mathbf{x}$ associated with individual policies in each period t. Standard parameteric distributions, such as multivariate Gaussian would be of little help due to both lack of flexibility and difficulty in interpreting e.g. the covariance matrix, whereas a non parametric technique such as a kernel density estimator will arguably provide even less insight. On the other hand an exponential family distribution (see Appendix B) with sufficient statistics roughly corresponding to the derived features used in models of claim frequency and severity can be readily understood.

Working with exponential families of distributions with user specified sufficient statistics is not (well) supported by available statistical software, if only because computing the log-partition function $A(\theta)$ quickly becomes intractable as the dimension of $\mathbf{x}$ increases. Such an approach might not have been very helpful in light these limitations, but it turns out that the *density ratio* of two exponential family distributions can be easily handled.

In particular it can be shown (see Appendix D) that if we label a sample of policies in force in one period as negative examples $y_{0i} = 0$, those in a subsequent period as positive examples $y_{1i} = 1$ and fit a logistic model to these data, then the following quantity can be interpreted as an estimator of the density ratio:

$$\frac{p(\mathbf{x}|\theta_1)}{p(\mathbf{x}|\theta_0)} \approx \frac{\hat{\pi}_0}{\hat{\pi}_1} \exp\left(\hat{\mathbf{w}}^T \phi(\mathbf{x}) + \hat{b}\right), \tag{36}$$

where $\hat{\mathbf{w}} \approx \theta_1 - \theta_0$ is the parameter vector of the logistic regression and $\hat{\pi}_0, \hat{\pi}_1$ are the counts of policies in force in respective periods. This is a very intuitive scheme and indeed its variant has been reported in the insurance context [Pol03].

We can naturally extend the idea by making some assumptions about the evolution of the parameters θ_t. If we impose a suitable prior on the quantity $\theta_t - \theta_0$ we arrive to a logistic regression model with a penalty terms of the form $\|\mathbf{w}_t\|$, controlling the magnitude of individual parameters. By considering priors on $\theta_{t+1} - \theta_t$ we obtain a dynamic GLM penalising the first order differences

$\|\mathbf{w}_{t+1} - \mathbf{w}_t\|$, as θ_0 terms cancel out.

4.3 Claim size and frequency

Automatic updating of technical cost models is perhaps the most direct application of dynamic GLMs. Insurers perform periodic recalibrations of the regression models underlying their premium rates, however this process this may not be responsive enough to act on or even detect rapid changes in the claims experience. To address this we can introduce time evolution of model parameters as described earlier, e.g:

$$\underset{\mathbf{w}_1,\ldots,\mathbf{w}_m}{\text{minimise}} \quad \sum_{t=1}^{m} \mathcal{L}(\mathbf{y}_t; X_t\mathbf{w}_t) + \lambda \sum_{t=1}^{m-2} \|\mathbf{w}_t - 2\mathbf{w}_{t+1} + \mathbf{w}_{t+2}\|_2^2, \tag{37}$$

where $\mathcal{L}(\mathbf{y}_t; X_t\mathbf{w}_t)$ is the log-likelihood function for e.g. the Poisson or gamma distributions and the regularisation term penalises departures from linearity in the estimated time trends. The model can also be easily extended to work directly with *smooth effects* rather than individual parameters.

One difficulty is that the new claim information would need to be sufficiently well developed before including in a dynamic GLM. The time required for this will depend on the class of business. Modifications of the basic model to permit inclusion of not fully developed data may be a topic for further investigation. Seasonality adjustments may also be needed.

Monitoring claims experience within this framework has the advantage that not only does it indicate when new experience differs from previous model results, it also provides new parameter estimates as a byproduct, which could in principle be used as an input into a pricing algorithm. In the case of statistical case estimates, it may be possible to alter the claim management "strategy", e.g. by reviewing the legal panel or changing staff allocations and validate the results.

Figure 3 shows a simple example of such a monitoring system. We examine a claim size model for a motor portfolio with the dataset containing 26,964 claims across 50 months. Initially the model structure was determined using standard tests and goodness of fit measures. A time interaction was introduced for the intercept and several of the important parameters. The light grey lines represent the rather noisy parameter values found from interacting the model parameter with the time variable (month). To overcome the inherent volatility we include a regularisation term as in (37). This acts to ensure the parameter estimates change smoothly over time. The coloured lines represent the estimates as new data become available, blue being the first estimate at $t = 25$ through to the orange at $t = 50$.

4.4 Competitive position

Insurer's competitive position in the market can change quickly and yet it is a key input into the pricing algorithm. Having obtained a sample of competitor quotes, quantities such as expected value or variance are not readily interpretable and it may be beneficial to move beyond the GLM methodology. One compelling alternative is to use quantile regression (see Appendix A) to obtain models for a range of market positions e.g. 10^{th},15^{th},25^{th} etc. percentiles directly. It will then be possible to evaluate the market position for a given policy by comparing the proposed rate against the estimated percentiles.

While in principle it may be possible to obtain an arbitrarily large sample of quotes, thus obviating the need for a dynamic model, there are usually considerable costs involved and some form of time aggregation may still be needed, e.g.:

$$\underset{\mathbf{w}}{\text{minimise}} \quad \sum_{t=1}^{m}\sum_{i=1}^{n} \rho_\tau(y_{ti} - \mathbf{x}_i^T\mathbf{w}_t) + \lambda \sum_{t=1}^{m-1} \|\mathbf{w}_{t+1} - \mathbf{w}_t\|_1 \tag{38}$$

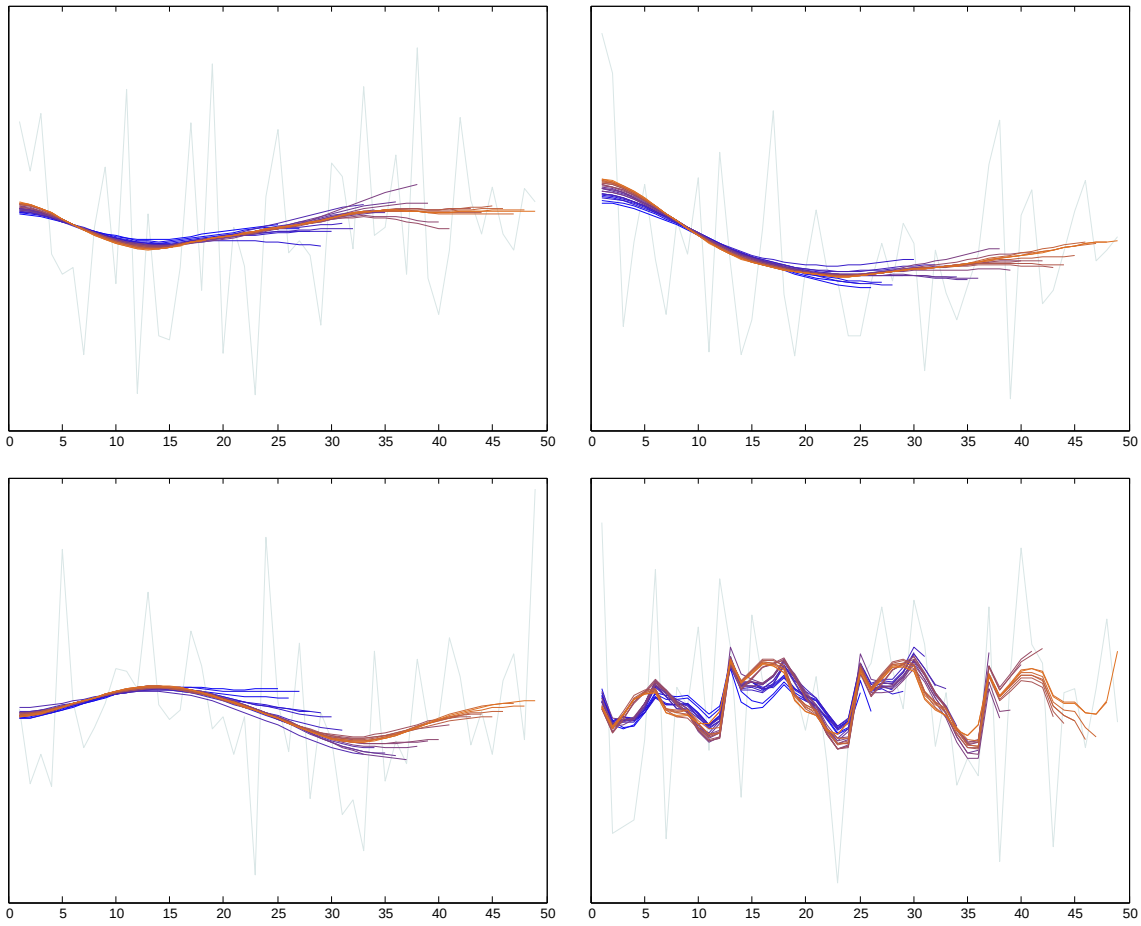

Figure 3: Sample output of a basic monitoring system. Additional details are in the main text. Each coloured line represents the model view of the historical values of the associated parameter. Colour gradient denotes the time at which the estimates are formed. The grey lines in the background show the parameter values for the models fit independently to each month's data. *Bottom right:* This is the intercept term, including a monthly seasonal adjustment

One also could apply quantile regression to determine if any significant risk factors are missing from an existing pricing model. This can be accomplished by fitting models using known risk factors to high and low quantiles of a single competitor's rates. Large differences indicate areas for investigation.

5 Algorithms

All of the problems described in this paper can be solved using standard algorithms [BV04] developed for convex optimisation. There are many high quality open source and commercial solvers, including SDPT3, SEDUMI, MOSEK and CVXOPT. As a rule these require problems to be converted to one of several standard forms (e.g. a second order cone program) but this step can be automated by using CVX, an excellent open source modelling system [GB08, GB10] for general convex optimization. It provides a convenient language for specifying convex problems and performs necessary mathematical steps to transform them into one of the forms that can be passed to a solver. It is often possible to work with sparse instances having millions of variables.

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 prob-

lems it may sometimes be necessary to implement a simple iterative reweighting scheme similar to Fisher scoring, such as in [LLAN06], with CVX in the inner loop. For very large problems decomposition methods are available, see [BPC+11] for an accessible introduction and examples.

Appendix

A Quantile regression

One interpretation of the least squares procedure (20) is that it estimates the conditional mean of y_i given the data vector $\mathbf{x}_i$. Regression with the asymmetric quantile [Koe05] loss function ρ_τ on the other hand results in estimates approximating the conditional τ-th quantile of the response variables y_i.

$$\mathcal{L}(y_1, \dots, y_n; X\mathbf{w}) = \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{w}^T\mathbf{x}_i), \quad \rho_\tau(u) = \begin{cases} \tau u, & u > 0 \\ -(1-\tau)u, & u \leq 0 \end{cases} \tag{39}$$

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 minimise $\frac{1}{2}\|\mathbf{y} - X\mathbf{w}\|_1$. We should note that quantile regression appears quite attractive for a number of practical applications (e.g. insurance) as it can provide a non-parametric estimate of the full conditional distribution of the dependent variable and can deal with such issues as concentration of probability mass at a certain point.

B Exponential families

Let us consider an exponential family distribution on y:

$$p(y|\mathbf{w}) = h_0(y) \exp\Big(\sum_{k=1}^{m} w_k \phi_k(y) - A(\mathbf{w}) \Big). \tag{40}$$

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, $\phi(y) = [\phi_1(y), \dots, \phi_k]^T$ is the vector of *sufficient statistics* and $A(\mathbf{w})$ is the logarithm of the normalizing constant or the *log-partition* function, namely:

$$A(\mathbf{w}) = \log \left(\int_{(y) \in y} \exp\Big(\sum_{k=1}^{m} w_k \phi_k(y) \Big) h_0(y)\, dy \right), \tag{41}$$

with summation replacing the integral for discrete distributions. It is easy to see that given an independent sample $(x_1, \dots, x_n)$ the maximum likelihood parameter estimation problem can be written as:

$$\underset{\mathbf{w}}{\text{minimise}} \quad \mathcal{L}(\mathbf{y}; \mathbf{w}) = \sum_{i=1}^{n} \Big(A(\mathbf{w}) - \sum_{k=1}^{m} w_k \phi_k(y_i) \Big). \tag{42}$$

It is a standard result that log-likelihood of distributions in the exponential families is concave in the natural parameters [Bro86, WJ08]. The solution is then characterised by the first order optimality conditions:

$$\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{y}; \mathbf{w}^*) = \sum_{i=1}^{n} \phi(y_i) - n \nabla_{\mathbf{w}} A(\mathbf{w}^*) = 0 \tag{43}$$

which imply that:

$$\nabla_{\mathbf{w}} A(\mathbf{w}^*) = \mathbb{E}_{\mathbf{w}^*}\phi(y) = \frac{1}{n}\sum_{i=1}^{n}\phi(y_i), \tag{44}$$

or that the method of maximum likelihood applied to an exponential family distribution seeks to find a parameter vector $\mathbf{w}^*$ such that the expected value of the sufficient statistics vector $\phi(y)$ under $p(y|\mathbf{w}^*)$ matches its sample average.

C Conditional exponential families

Below we discuss convexity properties of conditional exponential families, closely related to generalized linear models.

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

$$\begin{aligned} p(y, x|\mathbf{w}) &= h_0(y, x)\exp\Big(\sum_{k=1}^{m} w_k\phi_k(y, x) - A(\mathbf{w})\Big) \\ &= h_0(y, x)\frac{\exp\Big(\sum_{k=1}^{m} w_k\phi_k(y, x)\Big)}{\exp\big(A(\mathbf{w})\big)}. \end{aligned} \tag{45}$$

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, $\phi(y, x) = [\phi_1(y, x), \dots, \phi_k(y, x)]^T$ is the vector of *sufficient statistics* and $A(\mathbf{w})$ is the logarithm of the normalizing constant or the *log-partition* function, namely:

$$A(\mathbf{w}) = \log\Big(\int_{(y,x)\in\mathcal{Y}\times\mathcal{X}} \exp\Big(\sum_{k=1}^{m} w_k\phi_k(y, x)\Big)h_0(y, x)\,dx\,dy\Big), \tag{46}$$

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:

$$p(y|x, \mathbf{w}) = h_0(y, x)\exp\Big(\sum_{k=1}^{m} w_k\phi_k(y, x) - A(\mathbf{w}|x)\Big), \tag{47}$$

with conditional log-partition function given by:

$$A(\mathbf{w}|x) = \log\Big(\int_{y\in\mathcal{Y}} \exp\Big(\sum_{k=1}^{m} w_k\phi_k(y, x)\Big)h_0(y, x)\,dy\Big). \tag{48}$$

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 $i = 1, \dots, n$, the joint conditional probability can be written as:

$$\prod_{i=1}^{n} p(y_i|x_i, \mathbf{w}) = \prod_{i=1}^{n}\Big(h_0(y_i, x_i)\exp\Big(\sum_{k=1}^{m} w_k\phi_k(y_i, x_i) - A(\mathbf{w}|x_i)\Big)\Big), \tag{49}$$

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

$$\underset{\mathbf{w}}{\text{minimize}} \quad \sum_{i=1}^{n}\Big(A(\mathbf{w}|x_i) - \sum_{k=1}^{m} w_k\phi_k(y_i, x_i)\Big). \tag{50}$$

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 **??**).

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[6] referred to as a *conditional random field* [SM12] and is likely to prove quite fruitful in insurance applications.

To recover the (now classical) generalized linear models of Wedderburn and Nelder, consider the case when $\mathcal{X} = \mathbb{R}^k$, the carrier measure h_0 does not depend on $\mathbf{x}$ and with a particularly simple choice of sufficient statistics:

$$\phi_k(y, \mathbf{x}) = y x_k. \tag{51}$$

We can then rewrite (47) as a single parameter exponential family with respect to $\mathbf{w}^T\mathbf{x}$ in the following way:

$$\begin{aligned} p(y|\mathbf{x}, \mathbf{w}) &= p(y|\mathbf{w}^T\mathbf{x}) \\ &= h_0(y) \exp\Big(y\mathbf{w}^T\mathbf{x} - B(\mathbf{w}^T\mathbf{x})\Big), \end{aligned} \tag{52}$$

with the log partition function:

$$B(\theta) = \log\Big(\int_{y\in y} \exp(\theta y) h_0(y)\, dy\Big) \tag{53}$$

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

$$\underset{\mathbf{w}}{\text{minimize}} \quad \mathcal{L}(y; X\mathbf{w}) = \sum_{i=1}^{n} \big(B(\mathbf{w}^T\mathbf{x}_i) - y_i \mathbf{w}^T\mathbf{x}_i\big). \tag{54}$$

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

$$\mathbb{E}(y|\mathbf{x}, \mathbf{w}) = \nabla_\theta B(\mathbf{w}^T\mathbf{x}), \tag{55}$$

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 [Jør87]:

$$p(y|\mathbf{x}, \mathbf{w}, \lambda) = h_0(y, \lambda) \exp\Big(\lambda\big(y\mathbf{w}^T\mathbf{x} - B(\mathbf{w}^T\mathbf{x})\big)\Big), \tag{56}$$

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 $\mathbf{w}$ 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 [DGP97].

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.

[6] 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.

D Density ratios for exponential families

The use of logistic regression for direct density ratio estimation was proposed by Qin [Qin98] in the context of experiment evaluation. Consider a ratio of two exponential family distributions:

$$\begin{aligned}\frac{p(\mathbf{x}|\theta_1)}{p(\mathbf{x}|\theta_0)} &= \frac{\exp\left(\theta_1^T\phi(\mathbf{x}) - A(\theta_1)\right)}{\exp\left(\theta_0^T\phi(\mathbf{x}) - A(\theta_0)\right)} \\ &= \exp\left((\theta_1-\theta_0)^T\phi(\mathbf{x}) - A(\theta_1) + A(\theta_0)\right) \\ &= \exp(\mathbf{w}^T\phi(\mathbf{x}) + b)\end{aligned} \tag{57}$$

Rather than recovering the densities involved, it might be easier to estimate $\mathbf{w} = \theta_1 - \theta_0,\ b = A(\theta_0) - A(\theta_1)$ directly from data, ignoring the fact that both $\mathbf{w}$ and b are in fact both dependent on θ_0, θ_1. We attempt to reduce this problem to the standard logistic regression. Consider a categorical random variable y such that $p(y=k) = \pi_k$. We then assume the following form for $p(y|\mathbf{x})$ with an (unknown) parameter vector $\mathbf{w}$:

$$\begin{aligned}p(y=1|\mathbf{x}) &= \frac{\exp\left(\mathbf{w}^T\phi(\mathbf{x}) + b\right)}{1+\exp\left(\mathbf{w}^T\phi(\mathbf{x}) + b\right)} \\ p(y=0|\mathbf{x}) &= \frac{1}{1+\exp\left(\mathbf{w}^T\phi(\mathbf{x}) + b\right)}\end{aligned} \tag{58}$$

Having set $y = k$ for samples drawn from $p(\mathbf{x}|\theta_k)$ and given that there are n samples in total, we can obtain estimates $\hat{\mathbf{w}}, \hat{b}$ directly from data by minimising the corresponding negative log-likelihood, which coincides with logistic regression:

$$\begin{aligned}\mathcal{L}(\mathbf{y};\mathbf{w},b) &= -\sum_{i:\,y_i=1}\log\left(\frac{\exp(\mathbf{w}^T\phi(\mathbf{x}_i)+b)}{1+\exp(\mathbf{w}^T\phi(\mathbf{x}_i)+b)}\right) - \sum_{i:\,y_i=0}\log\left(\frac{1}{1+\exp(\mathbf{w}^T\phi(\mathbf{x}_i)+b)}\right) \\ &= \sum_{i=1}^{n}\Big(\log(1+\exp(\mathbf{w}_t^T\phi(\mathbf{x}_i)+b)) - y_i(\mathbf{w}_t^T\phi(\mathbf{x}_i+b))\Big).\end{aligned} \tag{59}$$

Then by the Bayes rule we can write:

$$\begin{aligned}\frac{p(\mathbf{x}|y=1)}{p(\mathbf{x}|y=0)} &= \frac{p(y=1|\mathbf{x})\pi_0}{p(y=0|\mathbf{x})\pi_1} \\ &\approx \frac{\hat{\pi}_0}{\hat{\pi}_1}\exp\left(\hat{\mathbf{w}}^T\phi(\mathbf{x}) + \hat{b}\right),\end{aligned} \tag{60}$$

which provides a direct estimate of the density ratio (57) after correcting for the difference in sample sizes.

References

[BG05] H Bühlmann and A. Gisler. *A Course in Credibility Theory and its Applications*. Springer, 2005.

[BHMM09] J. G. Berry, G. Hemming, G. Matov, and O. Morris. Report of the model validation and monitoring in personal lines pricing working party. In *GIRO Convention*, Edinburgh, 2009. Institute of Actuaries and Faculty of Actuaries.

[Boh99] G. Bohlmann. Ein Ausgleichungsproblem. *Nachrichten Gesellschaft Wissenschaften Gottingen. Math.-Phys. Klasse*, pages 260–271, 1899.

[BPC+11] 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.

[Bro86] L. Brown. *Fundamentals of Statistical exponential families with applications in statistical decision theory*. Institute of Mathematical Statistics, Hayward, CA, 1986.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Presss, 2004.

[CCFY86] 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.

[DGP97] D. K. Dey, A. E. Gelfand, and F. Peng. Overdispersed generalized linear models. *Journal of Statistical Planning and Inference*, 64:93–108, 1997.

[Fah92] 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.

[FK91] L. Fahrmeir and H. Kaufmann. On Kalman filtering, posterior mode estimation and Fisher scoring in dynamic exponential family regression. *Metrika*, 38:37–60, 1991.

[GA10] M. Grewal and A. Andrews. Applications of Kalman filtering in aerospace 1960 to present. *IEEE Control Systems Magazine*, pages 69–78, June 2010.

[GB08] M. Grant and S. Boyd. Graph implementations for nonsmooth convex programs. In *Recent Advances in Learning and Control*. 2008.

[GB10] M. Grant and S. Boyd. CVX: Matlab software for disciplined convex programming, version 1.21, September 2010.

[HP97] R. Hodrick and E. Prescott. Postwar U.S. business cycles: An empirical investigation. *Journal of Money, Credit, and Banking*, 29:1–16, 1997.

[HTF08] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer, 2nd edition, 2008.

[JG75] D. Jones and H. Gerber. Credibility formulas of the updating type. *Transactions of the Society of Actuaries*, 27:31–46, 1975.

[Jør87] B. Jørgensen. Exponential dispersion models (with discussion). *Journal of the Royal Statistics Society Series B*, 49:127–162, 1987.

[JZ83a] 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.

[JZ83b] P. De Jong and B. Zehnwirth. Credibility theory and the Kalman filter. *Insurance: Mathematics and Economics*, 2:281–286, 1983.

[Kal60] R. Kalman. A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82:35–45, 1960.

[KKBG09] S.J. Kim, K. Koh, S. Boyd, and D. Gorinevsky. ℓ_1-trend filtering. *SIAM Review*, 51:339–360, 2009.

[KNP94] R. Koenker, P. Ng, and S. Portnoy. Quantile smoothing splines. *Biometrika*, 81:673–80, 1994.

[Koe05] R. Koenker. *Quantile regression*. Cambridge University Press, 2005.

[LLAN06] S. Lee, H. Lee, P. Abeel, and A. Ng. Efficient ℓ_1-regularized logistic regression. In *AAAI*, 2006.

[Meh75] R. Mehra. Credibility theory and kalman filtering with extensions. Technical Report RM 75-64, International Institute for Applied Systems Analysis, Schloss Laxenburg, Austria, 1975.

[MS85] L.A. McGee and S.F. Schmidt. Discovery of the Kalman filter as a practical tool for aerospace and industry. Technical Report TM 86847, NASA, Moffett Field, California, 1985.

[Pol03] C. Pollack. Using non-parametric techniques to understand your data. In *XIVth General Insurance Seminar*, Canberra, 2003. Institute of Actuaries of Australia.

[Qin98] J. Qin. Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85:619–639, 1998.

[Roc70] R.T. Rockafellar. *Convex Analysis*. Princeton University Press, 1970.

[Sch78] D. Schuette. A linear programming approach to graduation. *Transactions of Society of Actuaries*, 30:407–445, 1978.

[Sea81] H. L. Seal. Graduation by piecewise cubic polynomials: A historical review. *Blätter der DGVFM*, 15:89–114, 1981.

[SM12] C. A. Sutton and A. McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4:267–373, 2012.

[Tay08] G. Taylor. Second-order Bayesian revision of a generalised linear model. *Scandinavian Actuarial Journal*, pages 202–242, 2008.

[Tay11] G. Taylor. Statistical basis for claims experience monitoring. *North American Actuarial Journal*, 15:535–551, 2011.

[TSR+05] 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.

[Ver93] R. Verall. A state space formulation of Whittaker graduation with existensions. *Insurance: Mathematics and Economics*, 13:1–14, 1993.

[Voh05] R. V. Vohra. *Advanced Mathematical Economics*. Routledge, London and New York, 2005.

[WBAW12] B. Wahlberg, S. Boyd, M. Annergren, and Y. Wang. An ADMM algorithm for a class of total variation regularized estimation problems. In *Proceedings IFAC Symposium on System Identification*, volume 16, 2012.

[Whi23] E. Whittaker. On a new method of graduation. *Proceedings of the Edinburgh Mathematical Society*, 41:63–75, 1923.

[WJ08] M. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2008.

[Wri05] M. Wright. The interior-point revolution in optimization: history, recent developments, and lasting consequences. *Bulletin of the American Mathematical Society*, 42:39–56, 2005.